Reported vs. Logged ChatGPT Use

How does ChatGPT usage found in anonymised logs compare to what users report in surveys

Name:	Quinten Voncken
Student number:	5168937
Programme:	MSc Management of Technology
Project period:	May – October 2025

Chair / 1st supervisor:	Prof. dr. ir. M. (Maarten) Kroesen
2nd supervisor:	Dr. S. (Savvas) Zannettou

Delft University of Technology
Faculty of Technology, Policy and Management

September 26, 2025

Preface

As I complete this thesis, I also close my time as a student at TU Delft. Returning to campus for the last few occasions felt like both a conclusion and a quiet farewell to a formative period. This project has been academic and personal in equal measure, shaped by many conversations, steady iteration, and the practical realities of doing research alongside recovery from a knee injury.

In early 2025 I began looking for a topic. The rapid rise of AI—and my own day-to-day use of generative tools—made questions about actual usage both timely and compelling. This thesis brings those interests together: it examines how self-reported use compares with behaviour measured from voluntary logs, and what that means for policy, governance, and everyday practice.

I am deeply grateful to **Dr. S. (Savvas) Zannettou**. From the very start he provided clear direction, consistent feedback, and room when recovery demanded time. He helped shape the structure of this thesis and what "good" should look like. Thank you for the guidance and for setting a high, humane standard.

My sincere thanks also go to **M. (Maarten) Kroesen**. When our original chair, **Marijn Janssen**, had to withdraw for personal reasons before the kick-off meeting, Maarten stepped in and took on the role from an early stage. I appreciate both the willingness to take over and the steady support throughout completion.

I wish **Marijn Janssen** all the best. I am grateful for his early engagement and hope everything worked out well.

Finally, thank you to everyone who participated in the survey and, especially, to those who donated usage data. All donated files were anonymised and handled under a privacy-by-design protocol in a restricted workspace. Friends and family also deserve heartfelt thanks—for patience, encouragement, and for the many small acts that make long projects possible.

Delft, September 26, 2025
Quinten Voncken

Executive Summary

Everyday ChatGPT use spans study and work, but surveys and usage logs capture different parts of that behaviour. Organisations and educators need a clear picture not just of how often people use ChatGPT, but also how people interact with it and for what tasks. This thesis contributes three things: a clean, like-for-like comparison of surveys and logs without linking individuals; a simple, auditable way to map prompts to tasks using example prototypes; and practical guidance for measuring and monitoring everyday use.

We ran a short survey and invited optional data donation via the platform's native export. We did not link individual survey answers to donated logs. Analyses therefore compare three independent groups: the full survey (S, N=93), a survey-donor subset (Sdon, n=24), and a logs-donor cohort (Llogs, n=24). The sample is male-leaning and younger (about two thirds aged 18–35). Results are descriptive; we do not make population estimates.

Survey items were designed in advance to mirror what we can derive from logs across four facets: intensity (sessions per week, sessions per day, minutes per session), timing (three broad dayparts on one time base), form (prompt-length bands), and portfolio (main task families and their subtasks, with the option to select more than one). In the logs, we mapped free-text prompts to the same task list with a compact codebook. We first defined a handful of example prompts ("prototypes") for each task, based on a quick scan of donated prompts and recent studies. Each prompt was matched to the closest prototypes with a simple similarity check; a small rule-based fallback handled borderline ties. Validation shows that task-family results are reliable for prevalence comparisons, while subtask results (Q12–Q17) are informative but not perfect and should be read as directional.

Because the cohorts are unpaired and small, we compare full distributions rather than pairs. We report differences in medians (for numeric measures) and differences in percentages (for categories), each with 95% confidence intervals. We also account for making several comparisons and keep the focus on effect sizes.

Two headline results follow. First, intensity aligns: typical-day activity and minutes per session overlap, and weekly frequency is only modestly higher among survey donors (about +2.5 sessions/week). Second, form and portfolio diverge. Survey donors more often describe paragraph-length prompts (about +54 percentage points), while logs contain many more one-liners (about −42 percentage points). The logs cohort also spans more task families (median 4 vs. 2), especially Coding and Language/translation.

Taken together, self-reports give a workable signal for how much while donation-based logs add detail on how people interact and for what. Short, one-line, iterative or technical exchanges are easy to miss in surveys, so using both sources together gives a more realistic picture for policy, training and procurement. Chapter 2 reviews prior work on measurement and use patterns, and Chapters 3–6 implement, analyse and report the comparisons [32, 31].

Nomenclature

Abbrev.	**Meaning**
AI	Artificial intelligence
LLM	Large language model
ECDF	Empirical cumulative distribution function (distribution plots)
HL	Hodges–Lehmann location shift (robust median difference)
MW–U	Wilcoxon–Mann–Whitney rank-sum test (two-sample, non-parametric)
KS	Kolmogorov–Smirnov two-sample test (distributional shift)
CI	Confidence interval (usually 95%)
IQR	Interquartile range (P25–P75)
SMD	Standardised mean difference (balance measure)
FDR	False discovery rate
BH	Benjamini–Hochberg procedure (FDR control)
Δp	Difference in proportions; effect reported in percentage points
pp	Percentage points
V (Cramér)	Cramér's V (strength of association in contingency tables)
OR	Odds ratio (model output)
Macro-/Micro-F1	F1-scores (example-based vs. label-based averaging)
TP/FP/FN	True/False Positives/Negatives (label-wise counts)
PII	Personally identifiable information
GDPR	General Data Protection Regulation (EU)
JSON	JavaScript Object Notation (ChatGPT export `conversation.json`)
UUID	Universally Unique ID (generated when missing `conversation_id`)
Q + A	Prompt–answer pair (single user→assistant exchange)
RQ / SQ	Research question / Sub-question
MOT	Management of Technology (programme context)
RMF	(NIST) AI Risk Management Framework
CV	Cross-validation (model tuning)
L_2	Ridge penalty in logistic regression
ρ	Spearman's rank correlation (monotone association; $[-1, 1]$)
η	Correlation ratio (share of variance for numeric vs categorical)
r	Rank-biserial correlation (two-sample rank effect; $[-1, 1]$)
δ	Cliff's delta (two-sample rank effect; $[-1, 1]$)
W1	1-Wasserstein (Earth-Mover's) distance (distributional shift; ≥ 0)
H	Hellinger distance for discrete mixes (0=identical, 1=maximally different)
q (FDR)	Target false-discovery-rate level in BH-FDR control (e.g., q=0.10)

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

Large language models (LLMs) such as ChatGPT have moved from pilots to everyday tools across study and work. Recent surveys report steady uptake among adults and teenagers, and rapid experimentation in workplaces across roles and sectors [9, 10, 30, 12]. As organisations formalise policies and training, they need to know not only whether people use these tools, but also how: how often, in which contexts, with what input styles, and for which tasks. These details matter for governance (policy and compliance), resource allocation (licensing and infrastructure), capability building (training and support), and impact assessment (productivity and quality).

Measuring real-world use is challenging. Self-reports scale well and provide essential context (plans, devices, roles), but decades of validation work show that reported behaviour often diverges from logged behaviour in level, timing, and form; short or highly interactive use is particularly easy to miss [32, 60]. Logs and platform exports provide time-resolved evidence about when and how people interact, and they can characterise inputs (for example, prompt length) that are hard to recall. By logs here we mean the chat history a user can export from ChatGPT. We use that user-controlled export as our source of interaction records (timestamps and messages). However, exports lack rich background variables, and access depends on user consent [31].

Combined with a short survey, this lets us compare what people say with what their exported history shows using group-level contrasts only (no one-to-one linkage).

A practical middle ground is user-centric data donation: participants export their own records and choose to upload them. Combined with a short survey, this lets us compare what people say and what their exported history shows using group-level contrasts only (no one-to-one linkage). To make comparisons interpretable, survey items are designed in advance to match simple export-derived frames: sessions per week and per day, minutes per session, three broad dayparts on one time base, word-count bands for typical prompt length, and a compact list of task families. With that alignment in place, we compare two groups (survey donors and logs donors) using unpaired, distributional contrasts.

Two gaps motivate the study. First, few papers place survey responses and user-controlled platform exports side by side on the same frames for everyday LLM use. Second, many descriptions emphasise how much people use LLMs but say less about how they interact

(input form) and what for (portfolio breadth and mix), even though these dimensions shape policy, training, and return on investment. This thesis addresses both gaps with a donation-based comparison on the same frames and with uncertainty reported throughout.

1.2 Research objective and questions

The objective is to develop and apply a transparent measurement design that (i) aligns survey frames with simple export analogues across four facets—intensity (sessions per week, sessions per day, minutes per session), timing (dayparts), input form (prompt-length bands), and task portfolio (main families and their subtasks); (ii) compares independent survey-donor and logs-donor cohorts on those indicators; (iii) validates a compact way to map free-text prompts to the same task list; and (iv) identifies where cohorts align or diverge and which survey subgroups sit higher or lower on the divergent components.

Research question

> *How does ChatGPT usage observed in anonymised logs compare with what users report in surveys?*

We address this question through three focused sub-questions.

SQ1. Which underlying dimensions structure self-reported usage, and how are demographics and contexts related to them?

Here we use the survey only (full N=93 where relevant) to map associations across items, reduce them to a small number of interpretable dimensions (for example, intensity of use, study/work orientation, task breadth), and relate those dimensions to plan, device, status and field.

SQ2. To what extent do survey constructs and their indicators correspond to log indicators, viewed in terms of distributions and prevalence?

We build like-for-like indicators on both sides (intensity, timing, input style, and task portfolios) and compare the survey-donor subset (n=24) with the logs-donor cohort (n=24) at the level of distributions and shares rather than single means.

SQ3. Which components differ most strongly between the survey-donor subset and the logs-donor cohort, and which survey subgroups are associated with higher or lower values on these components?

We rank gaps and use survey-only models to characterise which profiles and contexts align with higher or lower levels, without making causal claims.

Across all three, frames match the units and categories later defined in Chapter 4 to ensure comparability between self-report and logs [60]. Details on the specific statistics and interval conventions appear in section 5.4.

1.3 Linkage with the MOT programme

This work is a socio-technical measurement and governance study aligned with MOT's focus on managing innovation under real-world constraints. It (i) designs a simple measurement

system that compares two separate datasets at the group level (platform-native export, data minimisation, no one-to-one linkage), (ii) analyses technology adoption and use with methods suited to small, non-probability samples (estimation-first, effect sizes with intervals), and (iii) translates findings into actionable guidance for organisations (policy, training, monitoring). The general approach is to align frames between sources and use unpaired, distributional contrasts to compare independent cohorts. The emphasis on system design, stakeholder trust, and evidence-based decision-making reflects core MOT learning goals.

1.4 Academic and societal relevance

The thesis contributes a transparent, donation-based comparison of surveys and logs; aligns survey frames with traceable log analogues; and validates a compact, multi-label task taxonomy for everyday LLM use [32, 60, 61, 31]. Methods prioritise reproducibility and small-sample robustness.

For practice, results offer realistic baselines about how ChatGPT is used (not only whether). Surveys capture "how much" and profiles; exports add detail on input form and task breadth. The setup keeps handling simple, store as little as needed and avoid linking survey answers to donated files [20, 56], so comparison remains feasible for pilots in education and the workplace. Teams can use the same frames (weekly/daily rates, dayparts, prompt-length bands, compact task list) to monitor usage and target training.

1.5 Thesis structure

This report is structured as follows. Chapter 2 (Related Work) scopes sources and reviews survey-versus-log measurement, task portfolios for LLMs, and outcomes context, motivating a multi-aspect design that compares sources at the group level. Chapter 3 (Data Collection Methodology) covers population, recruitment, instrument, consent, and the donation funnel (export of `conversation.json`); cohorts are independent by design (no one-to-one linkage). Chapter 4 (Data Processing & Variable Derivation) details parsing of exports into Q–A units; survey-aligned derivations (weekly and daily rates, dayparts, prompt length); the hybrid task-coding pipeline; and validation. Chapter 5 (Statistical Analysis Plan) sets unpaired, distributional contrasts with an emphasis on effect sizes and intervals suited to small samples. Chapter 6 (Results) reports the survey landscape (SQ1), survey–log convergence (SQ2), and the largest gaps and subgroup patterns (SQ3), with sensitivity checks.

2

Related Work

This chapter situates the study in three strands of prior work and explains how each strand informs the design choices implemented in Chapters 3–6. First, we review evidence on measuring technology use with surveys versus digital traces, focusing on known gaps between self-reports and logged behaviour and on practices for aligning frames so that comparisons are interpretable [32, 60]. Second, we summarise what recent studies report about what people do with general-purpose generative models, the main task families and their subtasks, and why a multi-label view is necessary when uses blend [1, 18, 61]. Third, we cover outcomes that are often examined alongside usage (productivity, quality, creativity) to anchor where our results fit in the broader debate without claiming causal effects [12, 19, 53].

Throughout, we keep the frames consistent with the design constraints introduced later: donation-based logs with privacy safeguards and no one-to-one linkage to survey records (unpaired, distributional contrasts) [31]. Where operational or normative choices matter (for example, data minimisation and masking of obvious identifiers), we rely on institutional guidance and regulatory analyses. These grey sources justify handling choices; they are not used to establish prevalence [20, 56, 17]. Peer-reviewed evidence anchors measurement and usage patterns; grey literature supports the practical steps that make the study possible and responsible.

The chapter proceeds as follows. §2.1 specifies scope and source selection, including how recency and transparency thresholds differ across themes. §2.2 reviews measurement work on surveys versus logs and introduces the alignment logic we adopt in Chapters 4 and 5. §2.3 synthesises evidence on task portfolios and usage settings and motivates the multi-label approach used in our coding pipeline. §2.4 distils the implications for our design. §2.5 summarises gaps and bridges to the research question and analysis plan.

2.1 Scope and source selection

Our objective is pragmatic: assemble a recent, method-transparent evidence base that (i) sets credible expectations for survey versus log alignment, (ii) describes prevalent task families for general-purpose models, and (iii) situates common outcome claims without over-generalising beyond our convenience sample (Chapter 3). We screened English-language publications across peer-reviewed venues, working papers, official guidance and large-scale sector surveys, with inclusion criteria tuned per theme.

Table 2.1: Scope by theme: what we looked for and what counted as admissible evidence.

Theme	Target constructs	Admissible sources	Recency focus
Measurement (survey vs. logs)	Discrepancies; frame alignment; session/prompt units	Peer-reviewed; working papers with full methods	2020–2025 (plus foundational)
Data donation & ethics	Consent flow; minimisation; non-probability sampling	Peer-reviewed; official guidance and regulators	2019–2025
Task portfolios (multi-label)	Main families; blended intents; sub-task patterns	Peer-reviewed surveys and studies; method-documented sector surveys	2023–2025
Outcomes and impacts	Productivity, quality, creativity (descriptive)	Peer-reviewed or working papers; sector reports	2023–2025

Sources and coverage

We searched across Google Scholar (broad coverage and forward citations), Web of Science Core and Scopus (fieldable filters and metadata), SSRN and arXiv (fast access to emerging methods), and the public sites of regulators and official bodies (ICO, UK Government, FTC, OECD). Sector reports with documented methods (for example, sampling frame and instrument) were included for adoption context [12, 30]. Regulator and operational guidance informed handling choices (anonymisation and minimisation) rather than prevalence claims [20, 56, 17].

Time windows

For AI usage and task portfolios we prioritised 2023–2025 because practices and interfaces change rapidly [1, 18]. For data donation, recruitment and privacy handling we included work from roughly 2019 onward [31, 36, 42]. For measurement fundamentals (survey vs. traces and logging frames) we combined well-cited foundations with updates from 2020–2025 where constructs remained stable [32, 60].

Peer-reviewed articles and working papers provide the primary evidence on measurement, task portfolios and outcomes [32, 60, 1, 19]. Grey literature has a bounded role: official guidance and regulator reports for operational decisions about anonymisation and data minimisation [20, 56, 17], and sector surveys to contextualise adoption levels and perceived value at scale [12]. These sources are cited for procedures or context, not to adjudicate fine-grained prevalence.

Screening and inclusion

Screening proceeded in two passes. First, titles and abstracts were checked against theme-specific criteria (on-scope constructs and basic method transparency). Second, full texts were read to confirm methodological detail (construct definitions; sampling and measurement frames for surveys; operational definitions for logs). After the human read, we used Claude AI as a reading companion to generate a short point-form summary of each candidate and to flag any obvious mismatches with our scope, Claude was chosen for its smooth interaction and consistently clear summaries in spot checks. This step acted as a cross-check only; it did not perform retrieval, decide inclusion, or write the synthesis. Working papers were included when methods and analysis were documented well enough to support replication or auditing (for example, [60, 19]).

The themes in Table 2.1 mirror the operational choices reported in later chapters. Measurement work motivates the like-for-like frames used to compare self-reports to logs (weekly

rates, dayparts, prompt-length bands) and the emphasis on distributional contrasts rather than paired records [60, 32]. Donation and ethics sources underpin the tiered consent, data minimisation and masking used in Chapter 4 [31, 20, 56]. Portfolio studies and multi-label guidance explain why we classify prompts into overlapping families and summarise to donors with thresholds that aim for interpretability rather than exhaustiveness [1, 61]. Outcomes mappings are used for context in Chapter 6 without causal claims [19, 12].

2.2 Measuring generative AI use

This section sets our design choices against prior work on how AI use is measured. Large surveys have been the main way to quantify awareness and adoption in 2023–2025 because they scale across countries and sectors [9, 10, 30, 12]. At the same time, a decade of validation research shows that self-reports and digital traces can differ in both level and shape, sometimes substantially [32, 60, 45, 21]. A practical approach in the literature is to pair a short survey with a donation of platform logs and to compare independent cohorts on like-for-like frames. The aim in such designs is pragmatic: retain the reach and context of surveys and use logs to sharpen what people actually do.

Surveys are good at telling us who uses a tool and in what context (plan, device, role). That is why many widely cited indicators for adults, teens, students and workers are survey-based [9, 10, 4, 18, 24, 30]. The trade-off is well known: self-reports can over- or understate levels, compress distributions, and miss short or highly interactive use [32, 60, 21]. In short, surveys are essential for prevalence and profiles, but they are not a ground truth for intensity or prompt form.

Digital traces (client logs, server telemetry, exports) avoid recall error and offer time resolution. They capture *when* and *how often* well, and they can describe input form (for example, prompt length) that people struggle to recall [60]. But pure telemetry often lacks demographics, is hard to access ethically at scale, and linking traces to surveys raises consent, privacy and governance issues; linkage can also change who participates [47]. Designs that keep survey and trace collection close without one-to-one linkage are therefore attractive.

Data donation is a practical middle ground: participants export and donate their own records under clear conditions [31]. Donors are recruited differently than general survey samples, but ecological validity is good when platform-native exports are used [39, 14]. Willingness depends on trust, a clear purpose, modest incentives and privacy guarantees [49, 36, 3, 58]; recruitment channels also shape who reaches the upload page and who converts [37]. Reported implementations vary, but common elements include clear consent, use of platform-native exports and data-minimisation safeguards; specific file formats or pipelines differ across studies.

Three implications recur in the measurement literature. First, treat survey and log samples as independent cohorts and compare full distributions rather than paired records. Second, align frames so that survey items have direct log analogues in familiar units. Third, minimise stored content and avoid one-to-one linkage, following anonymisation guidance [20, 56, 17].

These implications carry into the results. The survey maps perceived intensity and task portfolios and provides profiles (plan, device, status, field). The logs add time-resolved

Table 2.3: Measurement approaches used in the literature and where our design sits

Approach	Captures well	Main trade-offs (typical)
Self-report surveys	Awareness, adoption, context	Scalable with rich profiles, but recall and social-desirability biases can distort intensity and form.
Platform logs or telemetry	Timing, frequency, form	High temporal fidelity, but often lacks demographics and requires strong governance or platform access.
User-based data donation	Naturalistic traces with consent	Uses platform-native exports; privacy-preserving by design; donor pools are selective; careful onboarding needed.

and content-aware views (for example, prompt length and task labels) across the full donated horizon. Where the literature warns about survey-trace mismatch, good practice is to report distributional contrasts with uncertainty rather than relying on single means [32, 60]. This estimation-first stance is applied consistently in Chapters 5–6.

2.3 Where and how LLMs are used

This section locates our study in the two settings where ChatGPT and related large language models (LLMs) are most visible today: higher education and the workplace. We summarise what recent studies say about adoption and typical tasks in each setting and explain how those patterns shaped the task taxonomy and comparisons used later in Chapters 4–6. We prioritise sources with transparent methods and recent field evidence.

2.3.1 Higher education

Student surveys and early campus studies report substantial but uneven adoption. Global and regional snapshots from 2023–2025 show sizeable shares of students who have tried or regularly use ChatGPT for study tasks, with variation by programme and language background [1, 4, 18, 24]. Broader polls point in the same direction outside higher education: both adults and teens report rising use for schoolwork between 2023 and 2025 [9, 10].

Across sources, three activity clusters recur: writing and communication (outlining, drafting, rephrasing), language support (grammar and style improvement, translation), and study help (summarising readings, explaining concepts, practice questions) [1, 4]. Creative ideation appears but is typically secondary. These clusters directly informed three of our main task families (Writing and communication; Language/translation; Study/exam) and our choice to allow multiple selections. Students often mix uses within a single session (for example, outline → translate → polish), which a forced single choice would miss.

Institutional policy shapes edge cases. Ambiguity about permitted assistance can channel use towards phrasing and summarisation; explicit prohibitions on full drafting reduce, but do not eliminate, that category [4, 18]. We therefore frame tasks at a functional level (what the tool was used for) rather than inferring policy compliance from text. This keeps labels close to user intent and avoids over-interpreting institutional context from logs

Table 2.4: Task families in prior work and examples frequently reported

Family	**Typical examples in studies and reports**
Writing and communication	Outlining slides or emails; drafting and redrafting text; tone and style adaptation; summarising meetings or documents.
Language/translation	Grammar and style improvement in a target language; translation of passages or full texts; audience-specific rewriting.
Coding/programming	Generating snippets; debugging; explaining code; converting between languages; writing tests.
Brainstorming or ideation	Idea generation for assignments, campaigns or features; creative prompts; role-play.
Study or exam support	Explaining difficult concepts; summarising readings; practice questions and quizzes.

alone.

2.3.2 The workplace

Adoption in firms reflects both top-down initiatives and bottom-up experimentation. Industry and policy reports emphasise rapid uptake since mid-2023, with heterogeneity across roles and sectors [12, 30]. Two patterns matter for our design.

First, use concentrates where LLMs have clear strengths: drafting and editing communications; summarising meetings or documents; generating or explaining code; and translation or audience-specific rewriting. Case studies and surveys highlight these families and note that some use occurs outside formal approvals, especially early on [44, 12]. To capture this breadth without presuming a single primary purpose, we use a multi-label task taxonomy and report portfolio breadth alongside prevalence.

Second, adoption is uneven [19]. Skill demands, data sensitivity and governance constraints make some teams faster adopters than others. Policy work warns that productivity gains can widen gaps if access and training are unequal [30]. This motivates our descriptive stance in Chapter 5. Contrasts are distributional rather than population estimates, and we avoid causal claims about work outcomes.

2.3.3 Implications for our design

Three implications follow.

1) Mixed use is the norm. We let respondents select multiple families and code logs with a multi-label pipeline (Chapter 4), reporting which families appear and how broad the portfolio is. This mirrors observed practice and avoids over-interpreting a single primary task [1, 12, 30].

2) Labels should describe what was done, not whether it was permitted. Prior work shows policy signals channel behaviour but do not fully determine it [4, 18]. We therefore keep functional labels and avoid treating compliance as a category.

3) Intensity and orientation are distinct. A person can be heavy or light in either setting

[30]. In Chapter 5 (SQ1) we therefore extract a separate study/work orientation alongside intensity and breadth and carry that structure into the cohort comparisons in Chapter 6.

2.4 Implications for This Study

This section summarises, in one place, the design commitments implied by the literature. Implementation details follow in Chapters 3–4 and the analysis plan in Chapter 5. The literature above points to three practical imperatives for studying everyday ChatGPT use: measure multiple facets of behaviour rather than one headline, align self-reports with traceable log frames, and protect participants' privacy so donation remains feasible and ethical. This section translates those points into concrete choices implemented in Chapters 3–6.

Work on digital behaviour shows that "use" is multi-dimensional: frequency, session length, time of day, input form and task portfolio are only loosely coupled [60]. Single items (for example, "How often do you use it?") tend to miss that structure and are sensitive to recall and interpretation effects [32]. We therefore treat intensity (sessions per week, sessions per day, minutes per session), timing (broad dayparts), prompt form (typical length) and task portfolio (main families and subtasks) as separate, interpretable facets. Each facet appears in the survey in plain terms and has a direct analogue in the export so the two sources can be compared later (Chapter 4).

Self-reports are indispensable for context (plans, devices, roles, attitudes), but validation work shows that reported behaviour and observed traces often diverge in predictable ways [32, 60]. Donation-based designs offer a workable compromise: a short survey provides reach and profiles; anonymised logs reduce recall error and sharpen the picture of what people did over time [31]. In our design, the survey supplies the frames and profiles; the logs supply distributional benchmarks in the same units. We do not treat logs as a perfect ground truth; they complement self-reports where recall is coarse.

People rarely use ChatGPT for a single purpose, and prompts often blend writing with brainstorming, coding with explanation, or translation with style improvement. This is a multi-label setting [61]. Our taxonomy therefore allows multiple task families per donor and multiple subtasks within a family. On the log side, free-text prompts are mapped to this taxonomy with a small, fixed codebook and a hybrid routing strategy (prototype matching in an embedding space with a minimalist judge for borderline cases), chosen for transparency, reproducibility and cost [40, 38]. Validation metrics set expectations about where labels are precise and where they are indicative (Chapter 4).

Donation only works if participants trust that their data remain anonymous and cannot be re-identified [49]. We therefore (i) use the platform's native export so donors see exactly what leaves their account [14]; (ii) minimise data on ingestion by parsing to question-and-answer units and masking obvious identifiers [20]; and (iii) avoid creating one-to-one links between survey records and donated files. Even hashed joins can be risky in small samples and are not anonymous in a strict sense [56, 17]. Instead, we compare independent cohorts (survey-donor and logs-donor) and frame all contrasts as unpaired, distributional comparisons (Chapters 3, 5).

Recruitment via a company intranet and social media yields a heterogeneous but selective sample that skews male and young, which is typical for networked recruitment and

donation studies [15, 39, 31]. Following guidance for non-probability designs, we report descriptive distributions and effect sizes with uncertainty rather than population estimates and highlight where composition may shape results [37, 42]. Where helpful, external benchmarks (for example, platform usage by age) provide context [35].

To keep survey and logs commensurate, we use broad, readable units: weekly and daily rates; minutes per session; dayparts on a single local time base; prompt-length bands in words; and a compact task taxonomy tied to explicit prototypes. Chapter 4 details how each survey frame (Q7–Q11; Q12–Q17) is mirrored on the log side and how donor-level behaviour is summarised robustly (for example, medians for skewed quantities and Wilson intervals for small-n proportions) [26]. Where parameters are design choices (for example, dominance and aggregation thresholds), we flag them and check sensitivity rather than treating them as standards.

Four commitments follow and structure the rest of the thesis:

1. measure multiple facets of use (intensity, timing, form and portfolio) rather than a single headline;
2. build survey frames that the export can mirror one for one, so comparisons use the same units;
3. treat tasks as multi-label with a compact, auditable taxonomy and report validation to calibrate trust; and
4. protect donors by design (data minimisation and no linkage), accepting unpaired, distributional contrasts as the trade-off.

Chapters 3 and 4 implement these choices. Chapter 5 translates them into tractable comparisons. Chapter 6 reports where survey and log distributions align and where they diverge.

2.5 Chapter Summary and Gaps

This chapter mapped the landscape around real-world LLM use and its measurement. We combined peer-reviewed work with high-quality grey literature to (i) clarify how LLM use is currently measured (self-report vs. logs or telemetry), (ii) characterise usage patterns (single-task versus multi-label and mixed workflows), and (iii) summarise commonly reported outcomes (productivity, quality, creativity). We emphasised recent studies on AI use (post-2020 for data-donation work) and retained older, foundational sources where core concepts and methods originated. Scope followed the two-pass screening described in §2.1, with English-language sources and theme-specific recency windows.

What is established

- Measurement: Most field studies rely on self-report (surveys or diaries). Telemetry or platform logs are rarer and often siloed. When both sources appear, they are typically used side by side without one-to-one linkage—surveys provide who/contexts; logs provide when/how—so the two are complementary rather than interchangeable.
- Usage patterns: Many studies reduce use to a single dominant task or frequency, even though day-to-day practice mixes drafting, reviewing, debugging, searching, ideation and transformation. Multi-label characterisations are less common.

Issue in prior work	Consequence for our study	Chapter
Triangulation gap (survey vs. logs)	Build survey constructs with direct log analogues; compare cohorts at an aggregate, privacy-preserving level (no one-to-one linkage).	Ch. 3, 4
Patterning gap (single-task framing)	Use a multi-label use-case taxonomy to capture combined workflows; report portfolio breadth alongside prevalence.	Ch. 4, 6
Outcome breadth gap	Use outcomes from prior work for context; avoid causal claims and keep our analyses descriptive.	Ch. 2, 6

Table 2.5: From related-work gaps to design choices in this thesis.

- Outcomes: Reported effects cluster around productivity and quality; creativity is measured less consistently. Many evaluations use constrained tasks with uncertain ecological validity for everyday work or study.
- Scope: Evidence often centres on general-purpose LLMs but within specific domains (for example, education or programming). Sampling is commonly convenience-based and skews younger and more male.

Where the gaps remain

- Triangulation gap: Few studies place survey responses and usage logs side by side in a way that lets claims be checked against observable behaviour.
- Patterning gap: Everyday LLM use is multi-aspect; single-purpose taxonomies understate combined workflows (for example, search → draft → revise).
- Outcome breadth gap: Productivity and quality dominate the discussion; creativity and other perceived outcomes are less systematically captured alongside them.

Table 2.5 summarises how our design choices (Ch. 3–7) address the above gaps. In short, we use a deliberately multi-aspect survey, aligned ex ante with available logs, and a multi-label use-case taxonomy so that we can analyse how people combine tasks and how perceived outcomes travel with those patterns. We focus on general LLM use (not vendor-specific workflows), and we keep donation and analysis anonymous at the individual level to encourage contribution while allowing aggregate cross-checks.

Our scope was English-language sources. For AI-use and data-donation studies we prioritised recency (post-2020) and retained older, foundational work for core concepts. Some operational assumptions necessarily draw on high-quality grey literature where peer-reviewed documentation is not yet available. Those sources inform procedures, not prevalence.

Guided by this synthesis, Chapter 3 details the survey instrument, the anonymised donation workflow and the task taxonomy. Chapter 4 shows how survey frames map to logs. Chapter 5 lays out the distributional comparisons we use at small n, and Chapter 6 reports where survey and log distributions align and where they diverge most.

3

Data Collection Methodology

This chapter describes the study population and recruitment, the survey instrument, the consent procedure and the data–donation funnel. The design enables a clean comparison between self–reports and log–derived measures while protecting participants' privacy.

3.1 Population and Recruitment

We recruited three cohorts: a full Survey Sample (S, $N = 93$), a Logs-Donor cohort (Llogs, $n = 24$) who uploaded exports, and a size-matched Survey-Donor subset (Sdon, $n = 24$) used for unpaired contrasts with Llogs. Figure 3.1 shows the funnel from 93 survey starts to 80 completes and 24 uploads; these counts anchor the cohort definitions used throughout.

The study used a convenience sample recruited through two channels: a company intranet post at a mid–sized software/media firm (about 100 employees across functions) and two Instagram Stories from the author that were reshared within the personal network. This is a non–probability sample whose composition is shaped by channel reach and network dynamics [15].

Two features of the respondent pool matter from the start. The sample is male–leaning: roughly two–thirds identified as men. The age distribution skews young: about two–thirds fall between 18 and 35 years old. These skews likely reflect the reach of the Instagram calls and the author's network; Instagram–based recruitment often concentrates reach in specific age segments [39]. External benchmarks corroborate that Instagram usage is higher among younger adults [35]. In total, the announcements are estimated to have reached about 2,000 unique viewers through direct followers and re–shares. Together with the intranet post, this produced a heterogeneous but selective group suited for descriptive, distributional contrasts rather than population estimates; this stance is consistent with recruitment and selection effects observed in data–donation studies and with non–probability survey guidance [37, 42].

This recruitment approach has two implications for the remainder of the thesis. First, because the channel mix and the observed skews (gender, age) reflect convenience sampling, all cohort contrasts are framed as descriptive. Second, the cohorts used later are independent by design: the Survey Sample (S) contains completed questionnaires; the Log–Donor Sample (Llogs) contains participants who uploaded logs; and the Survey–Donor Subsample (Sdon) is a size–matched subset of survey respondents used for unpaired,

Table 3.1: Cohorts and sample sizes used in the study.

Cohort	Definition	N
Survey Sample (S)	All respondents with usable survey data	93
Logs–Donor cohort (Llogs)	Donated `conversation.json`	24
Survey–Donor subset (Sdon)	Size–matched subset of S for Sdon–Llogs contrasts	24

distributional contrasts with Llogs. There is no one–to–one linkage between survey records and donated files, so all comparisons centre on distributional gaps between self–reported and log–derived measures—an approach that is common in donation–based designs when linkage is impractical or undesirable [31].

3.2 Survey Instrument

The survey was designed to compare self–reports with log–derived measures [32]. It was in English and took about 2–5 minutes to complete. Its structure mirrors the outcomes analysed later, but all analytic construction is deferred to Chapter 4. The opening block situates respondents with a compact profile—current status (student, employed, other), age band, gender, broad study/work field, ChatGPT plan, and primary device—and two context cues about institutional policy and whether use typically occurs in study or work settings. These variables provide the backdrop for later contrasts and allow us to describe cohort balance transparently. Table 3.2 lists all survey items with the short titles we use throughout the thesis. From here on we refer to questions as Q+title; for example, Q7 Sessions/week, Q10 Usage timing, Q11 Prompt length. We use this shorthand consistently in Chapter 4 Data Processing, Chapter 5 Statistical Plan and Chapter 6 Results so cross-references remain unambiguous.

A second block captures usage in frames that can be mirrored by logs. Respondents report sessions in the last seven days, sessions on a typical day, average minutes per session, the time of day when they use the tool most, and a typical prompt length. Categories are intentionally simple and interpretable: for timing, the day is collapsed into a few clear periods; for prompt length, bands are phrased in plain language (one short sentence, a short paragraph, multiple paragraphs, or "varies too much"). Chapter 4 explains how each of these frames is mapped to log–derived analogues over each donor's full donated horizon. Designing items so that survey frames can be mirrored by digital traces follows recommendations from validation studies that compare self–reports to logged behaviour [60].

The third block asks what respondents use ChatGPT for. A small set of main task families appears first (writing and communication, brainstorming/fun, coding, language/translation, study/exam, plus "other"), followed by concise subtask lists that remain visible regardless of the initial selections. If a respondent did not select the corresponding main family at Q12, they could indicate this explicitly in the follow–ups ("I did not choose this category"), which keeps the subtask items comparable without forcing ratings of irrelevant categories. Two short attitude items close the instrument: the perceived importance of ChatGPT and whether use would continue if access became paid–only. In the analysis, main families (Q12) provide the primary task contrast between Sdon and Llogs; subtask shares (Q13–Q17) are used as supporting detail. The task families reflect dominant use clusters reported in recent student surveys [1].

Table 3.2: Survey questions with shorthand titles used in the thesis

Q	Survey question	Short title
Q1	What is your gender?	Gender
Q2	Which age group do you belong to?	Age group
Q3	Which ChatGPT plan are you currently on?	Plan type
Q4	Which device do you use most often to access ChatGPT?	Device
Q5	What best describes your current status?	Status
Q6	What best describes your main field of study or work?	Field
Q7	How many separate ChatGPT sessions did you have in the last 7 days?	Sessions/week
Q8	On a typical day, how many ChatGPT sessions do you start?	Sessions/day
Q9	On average, how long does a single ChatGPT session last?	Session length
Q10	When do you most often use ChatGPT?	Usage timing
Q11	How long are your typical prompts?	Prompt length
Q12	What tasks do you usually use ChatGPT for?	Task families
Q13	Sub-tasks for “Writing and communication”	Writing subtasks
Q14	Sub-tasks for “Brainstorming / fun”	Brainstorming subtasks
Q15	Sub-tasks for “Coding / programming help”	Coding subtasks
Q16	Sub-tasks for “Language practice / translation”	Language subtasks
Q17	Sub-tasks for “Study revision / exam prep”	Study subtasks
Q18	In the last month, what share of your sessions were for study or work tasks?	Study/work share
Q19	How important is ChatGPT for completing your study or work tasks?	Importance
Q20	If ChatGPT became paid-only tomorrow, would you still use it?	Paid-only use

3.3 Consent Procedure and Participation Flow

Participation was voluntary and proceeded under tiered consent. On the survey landing page, potential respondents first read a short information sheet that explained the aim of the study, data handling, anonymity and the right to withdraw without consequences. Consent was indicated via an explicit yes/no choice before any questions (Appendix A contains the full consent form). Contact details for the author and supervisor were provided for questions. Only those who consented advanced to the questionnaire.

At the end of the survey, a thank–you page offered a separate and optional consent step for data donation. Respondents who agreed were redirected to a simple upload page with clear instructions for exporting and submitting their ChatGPT history (`conversation.json`). The voucher lottery applied only to this donation step to offset the extra effort without influencing survey answers, in line with evidence on motivations and willingness to donate [36]. Using a distinct, optional consent for donation follows recommended practice in data–donation designs [31], and reflects psychological evidence that trust, transparency and modest incentives are central drivers of donation behaviour [49].

Figure 3.1 summarises the participation flow from survey start to optional donation. These counts show where attrition occurred and define the three analytic cohorts used later. All contrasts are strictly unpaired (no one–to–one linkage) to minimise re–identification risk and preserve anonymity [20, 56, 17].

Recruitment occurred in three call–outs that captured most traffic: two Instagram story waves and one intranet announcement. Approximate survey starts attributed to these waves are reported with the funnel in Figure 3.1. These channels align with the observed skews (male–leaning and younger age bands) and contextualise the donation conversion.

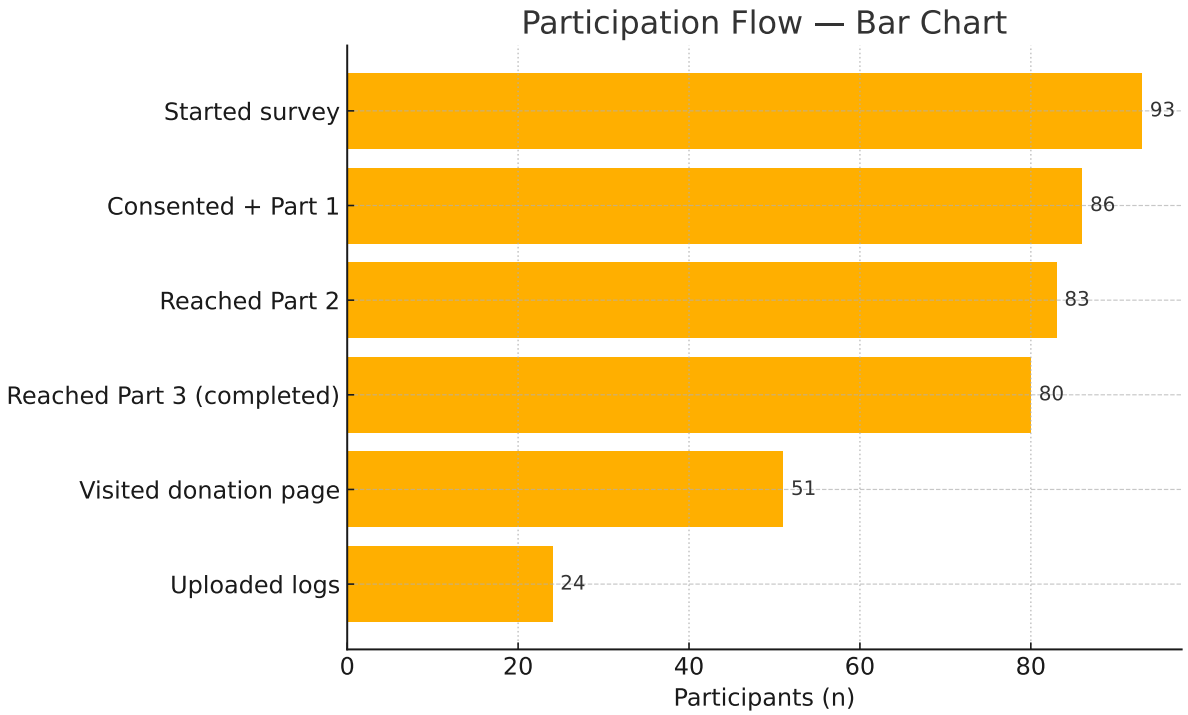

Figure 3.1: Participation flow from survey start to optional donation

3.4 Data–Donation Funnel

We used the platform–native ChatGPT export to keep the donation task simple and ecologically valid. Donors remained in full control: they saw exactly what is exported, decided whether to proceed, and uploaded the standard `conversation.json` file via a secure form, consistent with recent work on user–based data donation for generative AI [14].

Immediately after the survey, donors saw the official export instructions on the upload page:

1. Sign in at https://chat.openai.com.
2. Click the profile icon (top right) → *Settings*.
3. Open *Data Controls* and click *Export Data*.
4. Confirm export in the pop–up; an e-mail with a download link arrives within minutes.
5. Download the ZIP archive from the e-mail.
6. Unzip the archive; inside the sub-folder `ChatGPT`, locate `conversation.json`.
7. Upload only `conversation.json` through the study's secure upload form.

This hand–off keeps effort low and avoids custom software or browser extensions. The instructions appeared on the upload page, so donors did not need to navigate back and forth. Figure 3.1 provides context on how many respondents reached and completed the donation step.

Uploads were received in a restricted EU workspace. E-mail addresses for the voucher draw were collected separately and never entered the content pipeline. No personal

identifiers were collected and no one–to–one link was created between survey records and donated files. Upon receipt, files passed integrity checks (valid JSON and required fields) and were parsed into privacy–minimised question–answer (Q–A) records: one row per user→assistant exchange with only the fields needed downstream (see Chapter 4): pseudonymous `donor_id`, `conversation_id`, `turn_index`, prompt text, optional assistant text and timestamps. Personally identifiable strings (e-mail, phone, IBAN) were masked during parsing, in line with anonymisation guidance [20]. Prompt texts were retained under the same privacy safeguards; only fields needed downstream were stored. Time–based derivations use a fixed Europe/Amsterdam time base for interpretability in this population (details in Chapter 4).

4

Data Processing and Variable Derivation

This chapter turns the platform export (conversation.json) into analysis-ready question-and-answer (Q-A) records and derives log measures that exactly mirror the survey frames (Q7–Q11) plus task labels (Q12–Q17). We describe the pipeline Figure 4.1 and the retained fields (Table 4.1), define the rules and thresholds for sessions, dayparts and prompt-length bands (Table 4.2), and the validation checks that support reproducibility (Figure 4.2;Table 4.3). The outcome is a like-for-like set of log indicators that enable the unpaired comparisons reported in chapter 6.

4.1 Export and JSON Schema

We work with the official ChatGPT data export and ask donors to upload the platform-generated `conversation.json` [14]. Our goal is a compact table of question-and-answer (Q-A) records that mirrors the survey frames while storing as little as possible. Figure 4.1 shows the end-to-end pipeline from upload to analysis; Table 4.1 lists the fields we retain and what each is used for.

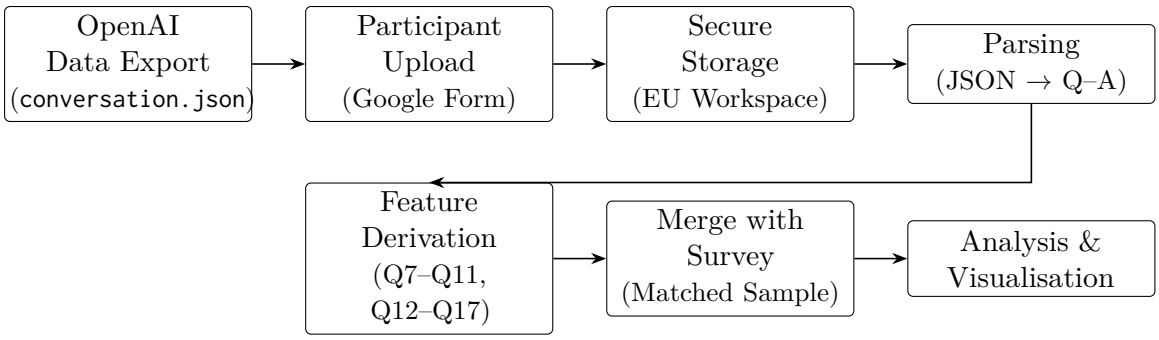

Figure 4.1: Data processing pipeline from export to matched analysis

From each exported conversation we traverse the message tree and collapse consecutive user→assistant turns into a single row. Treating a Q-A pair as the atomic unit avoids ambiguity in branched threads and provides a stable basis for all downstream measures. Where a session-level view is needed later (for example, to time an episode from first prompt to last reply), the `conversation_id` serves as the session key.

Data minimisation is deliberate. We keep a pseudonymous `donor_id` (generated per

upload, not a reversible hash), the `conversation_id`, a within-conversation `turn_index`, the user's `question_text` and the assistant's `answer_text` (optional), and timestamps for both. These fields are sufficient to construct session counts and durations (Q7–Q9), timing categories (Q10), prompt length (Q11), and task labels (Q12–Q17), without storing profile data or auxiliary metadata that are not needed for analysis [56].

Table 4.1: Parsed fields used in downstream derivations

Field	**Type**	**Used for**
`donor_id`	string	Grouping by donor; deletion schedule and audit.
`conversation_id`	string	Sessionisation (one conversation = one session).
`turn_index`	integer	Order of Q–A within a session; duration helper.
`question_time`	POSIX (Europe/Amsterdam)	Time windowing, rates, and dayparting (after local conversion).
`answer_time`	POSIX (Europe/Amsterdam)	Fallback for duration checks.
`question_text`	string	Prompt length (Q11) and task coding (Q12–Q17).
`answer_text`	string or null	Optional context for QA checks (not required for coding).

Before storage, obvious personally identifiable information is masked in both question and answer (for example, e-mail addresses, phone numbers, IBANs) [20]. We normalise whitespace and remove exact duplicate Q-A rows within a conversation; otherwise prompts are kept as written. We do not truncate or rewrite content. This keeps the textual signal intact for readers and for the classifier used later, while reducing accidental disclosure risk [20].

Timestamps are parsed from the export and stored in local Europe/Amsterdam time, and the same base is used for every time-dependent variable in this chapter. Using one local time zone avoids noise from unknown per-user settings and makes the daypart categories introduced in §4.2 directly comparable to the survey's options.

Basic integrity checks run on receipt: the JSON structure and required fields are validated. We maintain a small disposition ledger with file- and record-level reason codes so that exclusions are transparent, following general recommendations for data editing and transparent handling [57]. Uploads are handled in a restricted EU workspace; addresses collected for the voucher lottery are stored separately and never enter the content pipeline. Donated logs are deleted after thesis submission in October.

This normalisation yields a lean Q–A table used to construct log-side measures for the unpaired cohort comparisons in Chapter 6. To orient the reader, the next section opens with a compact overview of the five derived variables that mirror the survey frames (Table 4.2), followed by precise definitions for each of Q7–Q11.

4.2 Deriving Q7–Q11

As previewed at the end of §4.1, Table 4.2 summarises the five derived variables that mirror the survey items (Q7–Q11). We then define each quantity exactly and report the thresholds that are held fixed across sources, so that survey and logs remain directly comparable.

We first define what counts as a session. In the export a conversation_id captures a coherent threaded exchange. We therefore treat each unique conversation_id as a session in the primary specification. When users return to the same thread after a prolonged break, the behaviour resembles a new episode. To reflect that, we split a conversation into a new session after 30 minutes of inactivity between consecutive Q-A turns. A 30-minute inactivity cut-off is a pragmatic convention in digital-trace work to approximate natural breaks without over-fragmentation [48]. This behaviour-based rule provides a stable unit for rates, durations, and timing.

Q7. Weekly sessions

We count sessions over the donor's entire horizon and normalise to a weekly rate (sessions per seven calendar days). Normalisation avoids the artefact that donors with longer horizons would otherwise appear more active simply because more days are observed, and it aligns the trace window to the survey's weekly frame [60]. The resulting rates are banded into the same categories as the survey: 0, 1–2, 3–5, 6–10, >10.

Q8. Typical day

We mirror the everyday rhythm respondents report. We compute a per-day session count across the full horizon. Days with zero sessions are retained, except that we drop zero runs strictly longer than five days (`ZERO_STREAK_MAX`=5) to prevent extended off-study periods (holidays, outages) from dominating the statistic. This is a documented design choice; it follows robust-summary principles (reduce undue influence of long zero runs) and is paired with sensitivity checks reported alongside results. For the donor-level summary we use the median sessions per day over the kept days, a robust location estimator recommended in exploratory data analysis when distributions are skewed or zero-inflated [27]. Bands map directly to the survey categories: 0; 1; 2–3; 4–5; 6+.

Q9. Session length

For each session we measure elapsed time from the first user prompt to the last assistant reply. This captures the episode users actually experience, not just isolated turns. When timestamps are incomplete we apply a small, conservative proxy (baseline 3 minutes plus 3 minutes per additional Q-A turn) and bound durations to $[1, 480]$ minutes to avoid pathological values. Both choices are sanity checks grounded in robust EDA practice for handling outliers and partial information [27]. Donor-level means are then mapped to the Q9 minute bands: < 5, 5–15, 15–30, 30–60, > 60.

Q10. Time of day

Each session start is classified, in Europe/Amsterdam time, into one of three broad dayparts: During work/study hours (Monday–Friday, 09:00–18:00), Evenings (18:00–03:00, crossing midnight), and Other (the remaining hours and weekends outside those windows). These boundaries follow typical diurnal patterns reported for smartphone and computing use in comparable populations [22]. At donor level we assign one final label by dominance:

Table 4.2: Overview of derived variables for Q7–Q11

Variable	Unit	Operational definition (concise)
Q7. Weekly sessions	count/7d	Sessions per donor, normalised to a weekly rate; session = `conversation_id` split by 30 min inactivity. Bands: 0, 1–2, 3–5, 6–10, >10.
Q8. Typical day	mean/day	Sessions per calendar day over the full donated horizon; zero days kept; drop only 0-streaks >5 days. Bands: 0, 1, 2–3, 4–5, 6+.
Q9. Session length	minutes	First prompt to last assistant reply; proxy if missing times (3 + 3 min/extra Q-A); bound [1,480]; donor mean mapped to bands.
Q10. Time of day	category	Dominant daypart of session starts: Work/study (Mon–Fri 09:00–18:00), Evenings (18:00–03:00), Other; donor label via dominance ≥ 0.33, else "Anytime".
Q11. Prompt length	category	Median words per prompt; bands: ≤ 20, 21–60, > 60; "Varies too much" if no band ≥ 0.33.

if at least a third of a donor's session starts fall in a bucket, that bucket is the donor's timing category; otherwise the label is "Anytime throughout the day". The one-third dominance rule reflects typical use while keeping labels interpretable; in multi-label classifiers such thresholds are commonly tuned for interpretability rather than raw accuracy [61]. Using session starts prevents long sessions from overweighting a period, broad buckets reduce boundary noise, and a single Amsterdam time base makes interpretation straightforward for this population.

Q11. Prompt length

We count words per user prompt as written (after light whitespace normalisation) and take the donor's median as a robust summary of "typical", again to reduce sensitivity to outliers or a few very long prompts [27]. Medians are mapped to the survey's bands set a priori: one short sentence (≤ 20 words), a short paragraph (21–60), or multiple paragraphs (> 60). If no single band accounts for at least one third of a donor's prompts, we assign "Varies too much to say". These bands replicate the survey frames for direct comparability rather than claiming linguistic norms.

Together, these constructions prioritise interpretability and comparability. They keep the units readers expect from the survey, avoid fragile parameter choices, and make clear why each choice was made. Where values are design choices rather than standards (for example, 30-minute inactivity, 0.33 dominance, 5-day zero-streak), we flag them explicitly and check them in sensitivity analyses. The next section turns to task coding (Q12–Q17), where free-text prompts are aligned with the survey's task taxonomy.

4.3 Task Coding for Q12–Q17

To compare tasks fairly across survey and logs, free-text prompts must be mapped to the same categories respondents saw. We therefore label each user prompt against the survey's task taxonomy (Q12 main families; Q13–Q17 subtasks). Labelling happens at prompt level and is summarised to the donor level only afterwards. Codebooks with the prototype phrases that define each label are fixed ahead of time and listed in Appendix C. Keeping these prototypes explicit makes the measurement target transparent and supports reproducible auditing [38]. The resulting label space is multi-label by design so that blended intents can be represented without forced choice [61].

Prototypes: provenance and refinement

Prototypes were seeded directly from the survey instrument (the answer options for Q12–Q17) and a first pass over a small, stratified sample of prompts. We then refined them in three short steps using the validation workflow described in §4.4: (i) expand near-synonyms that appeared in correct matches; (ii) inspect borderline errors and add 1–3 disambiguating phrases for the most confused label pairs; (iii) freeze the codebook before the final pass. This keeps the codebook compact and auditable (see Table C.1 in Appendix C), while tuning it just enough to the data without overfitting.

Hybrid pipeline and design rationale

At a high level the pipeline is hybrid. We route most prompts by semantic similarity in an embedding space (vector representations of sentences) and resolve genuinely borderline cases with a compact LLM judge under tight output constraints. Concretely, we use text-embedding-3-small to embed the prompt and the prototypes for each family. This model provides stable sentence-level representations suited to paraphrase detection and prototype matching at low cost [40]. For small, fixed codebooks, it separates broad intents (for example, Writing vs. Coding) reliably and captures near-synonyms without overfitting to style. Cosine similarity then measures how close the prompt is to each label's prototypes. A label is accepted when it is clearly similar in absolute terms (operating point around 0.28 on a 0–1 cosine scale) or when it is essentially tied with the best match (within about 0.06 of the top score). This simple rule captures paraphrases without forcing a single winner when a prompt blends intents.

Why this structure and these thresholds? Three constraints guided the choice:

1. Valid labels at family and subtask level. Embeddings handle most prompts reliably for a small, fixed codebook; a judged fallback is only needed near boundaries where intents blend.
2. Reproducible and auditable outputs. The judge is constrained to return only label indices (JSON-only, temperature = 0), which keeps the step deterministic and cheap.
3. Runtime budget. Thresholds were tuned against the validation setup in §4.4 so that the full Q12–Q17 pass completes within the set runtime, while preserving micro-F1 at the family level. Practically, the operating point (0.28) and the narrow ambiguity band keep judged fallbacks to a small fraction of prompts; `REFINE_MAX_CALLS`= 800 caps worst-case calls per run.

Routing and aggregation

Ambiguous ties occur between neighbouring labels (for example, "summarise this article and translate the abstract"). In those cases we call the LLM judge (`gpt-5-mini`) that receives the Q-A pair and returns only the label indices. We invoke this step sparingly: when the top similarity lands in a narrow ambiguity band (roughly 0.345–0.355) or when the top two labels are separated by at most 0.010 [11].

Two policies keep labels readable. First, the pipeline is multi-label at prompt level: if a request genuinely combines tasks, it can carry more than one label [61]. Second, the fallback *Other* is exclusive and appears only when no concrete label clears the operating point; we never combine *Other* with concrete labels. This avoids double-counting and keeps *Other* a true residual. Subtasks are considered only for donors whose logs show the corresponding Q12 family. This hierarchical gating improves efficiency and label quality [51].

After prompt-level labelling, we summarise to the donor level across the full donated horizon so that the log side mirrors what the survey calls "typical use". A Q12 family is counted for a donor only when it appears with enough support to be plausibly typical: at least five prompts and at least ten percent of that donor's prompts ($n \geq 5$ and share $\geq 10\%$). Subtasks (Q13–Q17) use lighter thresholds, at least three prompts and at least ten percent ($n \geq 3$, share $\geq 10\%$), and are gated on their parent. The aim is a donor-level summary that reflects typical use rather than one-off occurrences. A 10% share, paired with $n \geq 5$ (Q12) or $n \geq 3$ (Q13–Q17), (i) scales with a donor's horizon, (ii) prevents a handful of prompts from very active donors from dominating, and (iii) keeps person-level labels interpretable. In small-sample checks tied to the validation workflow in §4.4, thresholds between 5–15% led to similar family-level conclusions; 10% offered the best trade-off between stability, readability and keeping the judged fallback small.

Inputs are the user prompts as written (after the basic redaction and normalisation described in §4.1). We do not alter content for classification. The outputs we carry forward are simple: for each donor, the set of Q12 families and gated subtasks that pass threshold, plus per-label counts and shares used in Chapter 6. Validation of this pipeline is reported next in §4.4.

Parameter values

The embedding step returns a cosine similarity between a prompt and each label's prototypes. The parameters below control when a label is accepted on similarity alone, when a borderline case is escalated to the judge, and how we keep runtime bounded and outputs deterministic (see §4.4 for the tuning logic).

- `ABS_DREMPEL`= 0.28 — minimum absolute similarity to accept a label outright; guards against spurious matches.
- `REL_MARGIN`= 0.06 — tie zone: if the runner-up is within this margin of the top score, we allow both (multi-label) or escalate.
- `BAND_LOW`= 0.345 & `BAND_HIGH`= 0.355 — narrow ambiguity band around the operating point; hits in this band trigger the judge.
- `TIE_GAP`= 0.010 — if the top two labels are closer than this gap, treat as a genuine tie (again, candidate for the judge).

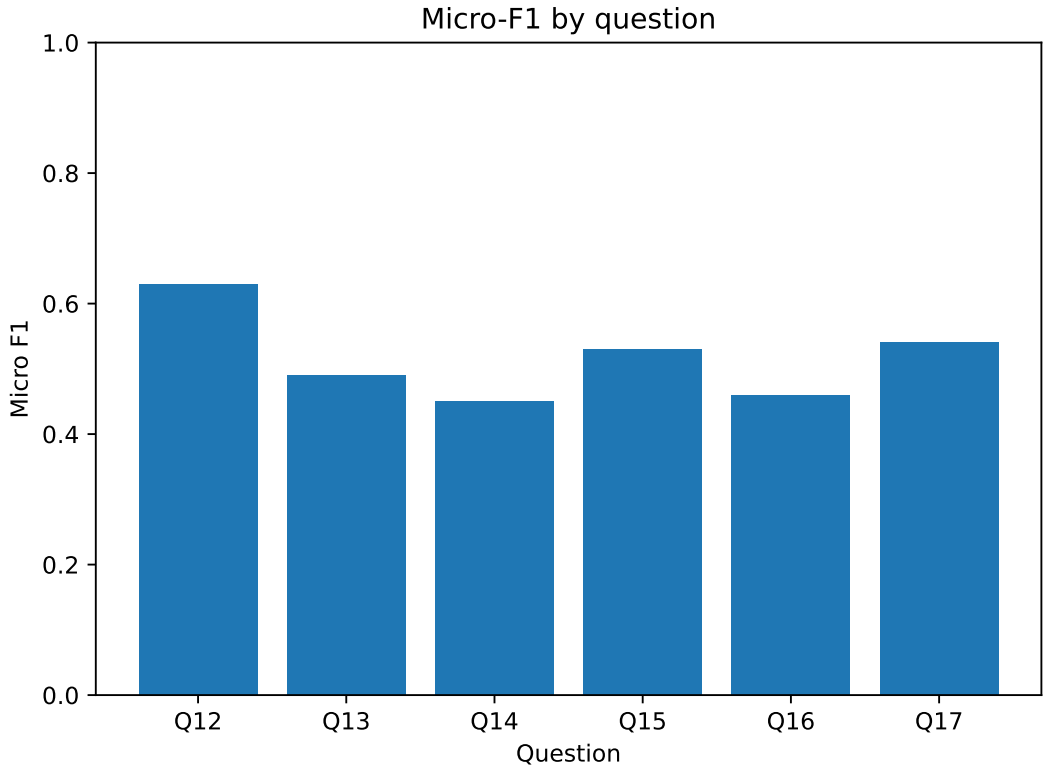

Figure 4.2: Micro-F1 by question (Q12–Q17).

- `MIN_TOP_FALLBACK`= 0.26 — if even the best similarity is below this floor, route to *Other* (we never combine Other with concrete labels).
- `REFINE_MAX_CALLS`= 800 per run; judge `gpt-5-mini`, temperature 0 — engineering caps that keep the end-to-end pass fast and reproducible.

4.4 Validation

Before interpreting any task distributions, we validate the coding pipeline described in §4.3. The aim is practical: to show how well the pipeline recovers the survey's task taxonomy from free-text prompts, where it performs strongly, and where outputs should be treated as indicative. Validation is conducted on six independent samples, one for each family of labels (Q12–Q17), with $n = 100$ Q-A pairs per family. Samples are stratified across donors with a fixed random seed. A single human rater assigned gold labels while blinded to model predictions and donor identity. Cohen's κ is not reported because only one rater produced the gold set (κ quantifies agreement between *two or more* raters); where κ is planned, recommended sample sizes depend on expected agreement and prevalence and can be substantive [8]. Agreement is reported with metrics that are familiar and interpretable in multilabel settings, with micro-F1 as the primary summary, weighted macro-F1, macro recall, and any-overlap accuracy, following common practice in multi-label evaluation [61]. For any-overlap accuracy—counting an example as correct if any true label is returned—we report Wilson 95% confidence intervals [26]. Given class imbalance in some families, these summaries prioritise precision/recall trade-offs over ROC-based summaries [43].

Question	N	Overlap acc	Overlap CI lo	Overlap CI hi	Macro F1	Macro Rec	Micro F1
Q12	100	0.63	0.53	0.72	0.65	0.75	0.63
Q13	100	0.49	0.39	0.59	0.55	0.67	0.49
Q14	100	0.45	0.36	0.55	0.53	0.68	0.45
Q15	100	0.53	0.43	0.62	0.62	0.71	0.53
Q16	100	0.46	0.37	0.56	0.55	0.68	0.46
Q17	100	0.54	0.44	0.63	0.59	0.86	0.54

Table 4.3: Validation summary per question (Q12–Q17).

Table 4.4: Code sets for Q12–Q17.

Code	**Q12: Main tasks**	**Q13: Writing subtasks**	**Q14: Brainstorming/fun subtasks**	**Q15: Coding subtasks**	**Q16: Language/-translation subtasks**	**Q17: Study/exam subtasks**
WRI	Writing & communication	Outlining ideas or slides	Academic or research topics	Generating new code snippets	Translating full texts between languages	Summarising lecture notes or readings
BRA	Brainstorming / fun	Drafting full text	Business or marketing concepts	Debugging existing code	Improving grammar or style in a target language	Generating practice questions or quizzes
COD	Coding / programming	Proof-reading / tone adjustment	Creative role-play, jokes, stories	Explaining code / concepts	Vocabulary drills or word lists	Explaining difficult concepts in simple terms
LAN	Language / translation	Summarising sources or meeting notes	Hypothetical "what-if" scenarios	Converting code between languages	Conversational practice / dialogue role-play	Reviewing flashcards / key terms
STU	Study / exam	Adjusting style for different audiences	Recommendations (books, movies, music)	Writing unit tests	Pronunciation or phonetic guidance	—
TRI	—	—	Trivia & general knowledge	—	—	—
OTH	Other	Other	Other	Other	Other	Other

Per-label diagnostics clarify where errors arise. For Q12 (the main families; Figure 4.3), Writing & communication and Study/exam show relatively high F1 (0.70 each), with precision/recall of 0.80/0.62 (support = 13) and 0.94/0.56 (support = 27). Language/-translation is solid (F1 = 0.62), while Brainstorming/fun trades perfect recall for many false positives (precision = 0.23, recall = 1.00, F1 = 0.38, support = 3), reflecting loosely phrased "ideas" prompts that overlap with adjacent categories. The residual *Other* is precise (precision = 0.76) but conservative on recall (0.56; F1 = 0.64; support = 45), which means some genuine uses remain in concrete labels rather than being swept into *Other*.

A representative subtask family shows the same pattern of strengths and near misses. For study/exam subtasks (Q17), Explaining difficult concepts in simple terms is strong (precision = 0.74, recall = 0.91, F1 = 0.82), while very small-support subtasks such as Summarising lecture notes or readings, Generating practice questions or quizzes, and Reviewing flashcards / key terms yield low and unstable F1 (supports = 2–3). The *Other* subtask category again shows high precision with modest recall (F1 = 0.55, support = 71), which is desirable for a residual class but implies that observed subtask shares in Chapter 6 should be read as lower bounds for concrete subtasks. See Figure 4.4 for the per-label F1

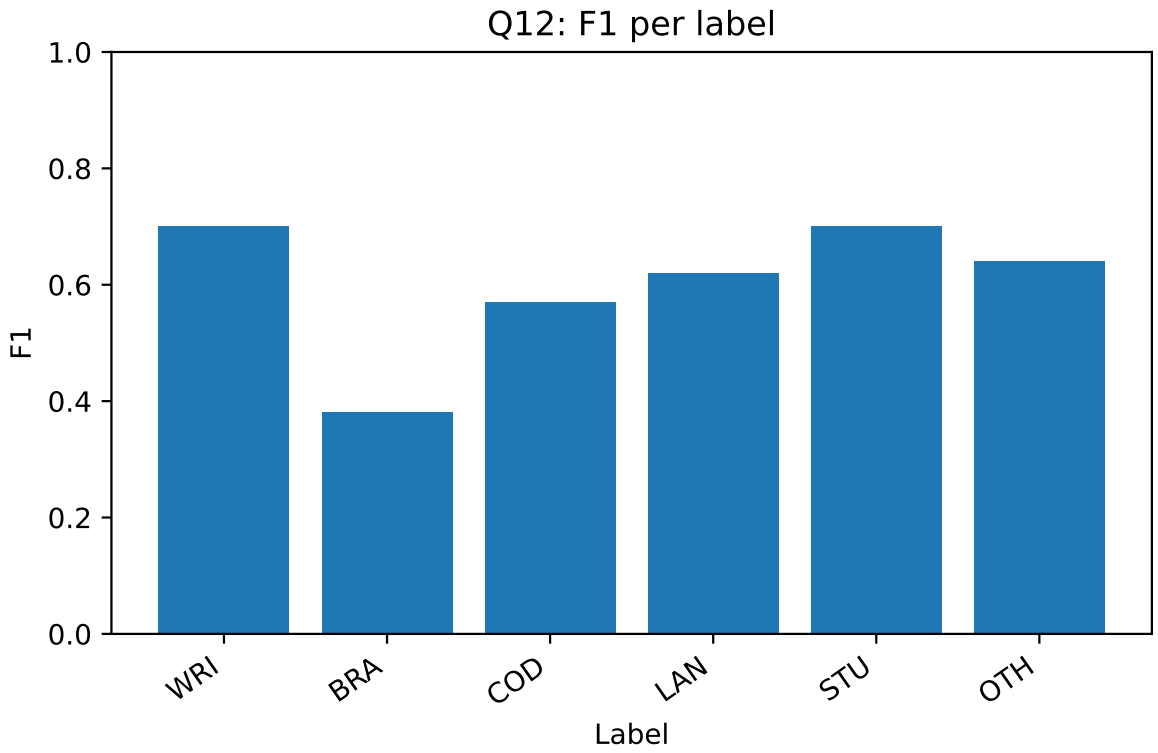

Figure 4.3: Q12 per-label F1.

view (codes per Table 4.4).

These diagnostics guide interpretation downstream. At the family level (Q12), micro-F1 = 0.63 and balanced recall = 0.75 support using Q12 shares for prevalence contrasts. At the subtask level (Q13–Q17), micro-F1 in the 0.45–0.54 band supports pattern-level comparisons rather than precise prevalence claims. Where subtask labels have very small support, variability is expected and documented rather than over-interpreted. All codebooks (prototype lists) are fixed before inference and reproduced in Appendix C, in line with reproducibility guidance [38]. The complete validation outputs accompany this chapter: the per-question master table appears here (Table 4.3); all per-label metrics tables for Q12–Q17 (TP/FP/FN, precision/recall/F1, support) are collected in Appendix D; the full set of per-label F1 panels and TP/FP/FN stacks is placed in Appendix E.

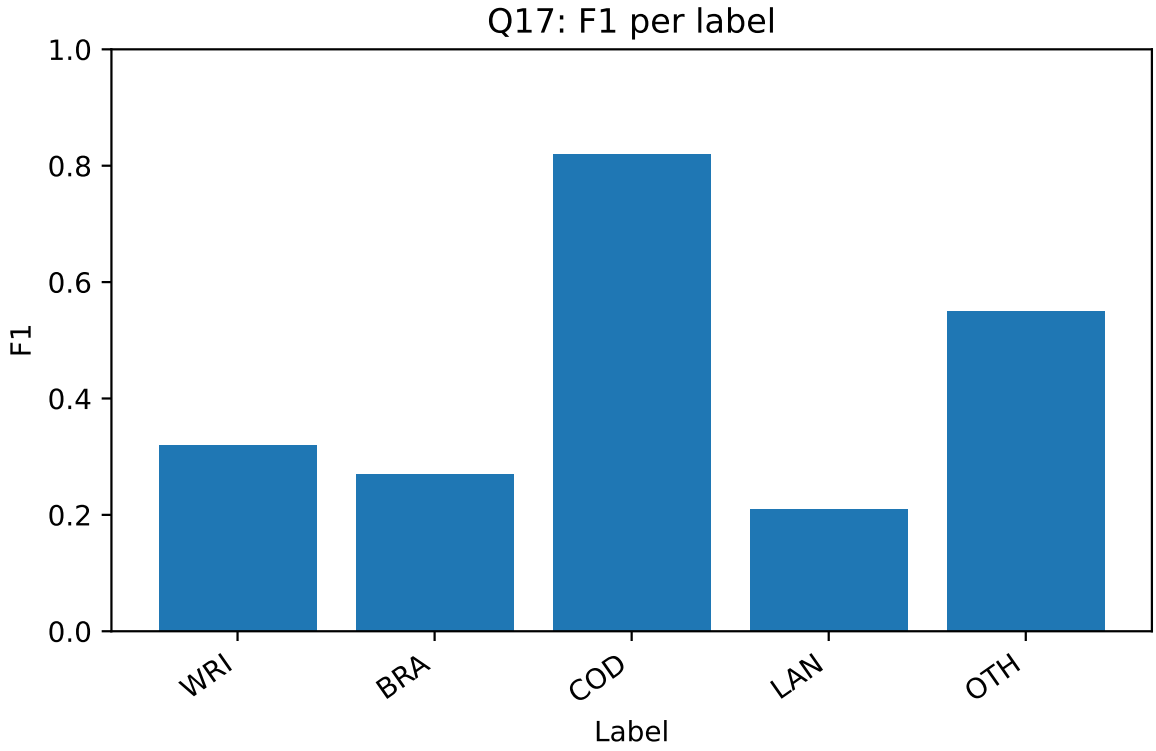

Figure 4.4: Q17 per-label F1.

5

Statistical Analysis Plan

This chapter sets out what we compare, how we compare it, and why these choices fit the data. Survey answers reflect perceived use; donated logs (processed in Chapter 4) reflect observed use. Because we do not link individuals one-to-one, we work with two independent cohorts: the survey-donor subset (Sdon, n=24) and the logs donors (Llogs, n=24). All contrasts therefore compare full distributions across independent groups, which is standard in donation-based designs without record linkage [31]. Given the small n, we focus on effect sizes with confidence intervals rather than on pass/fail significance tests [59]. Reporting conventions appear in section 5.4.

5.1 SQ1: Structure of Self-Reported Use

SQ1 describes the internal structure of the survey and who scores high or low on the resulting dimensions. We use the full survey (N=93) to learn the structure and apply the same scoring to Sdon (n=24) so later contrasts use identical constructs. The goal is twofold: (i) a clear picture of how respondents say they use the tool, and (ii) a small set of well-behaved variables that Chapter 6 can reference without re-explaining method.

Usage items (Q7–Q9) are ordered bands. Where a numeric summary helps, we map to band midpoints in their native units: Q7 as sessions in the last seven days; Q8 as median sessions per day; Q9 in minutes per session. Q11 Prompt length stays as an ordered set of three bands plus Varies too much to say. Multi-select items (Q12–Q17) become one binary indicator per option; the informational response I did not choose . . . is kept for transparency but does not add to counts. From these binaries we form breadth measures: the number of Q12 families a respondent selects (task breadth) and, within families, the number of subtasks selected. A simple intensity index, usage_index, is the unit-weighted mean of the z-scores of $\mathrm{Q7}_{\mathrm{mid}}$, $\mathrm{Q8}_{\mathrm{day,\ mid}}$ and $\mathrm{Q9}_{\mathrm{mid}}$. We use unit weights on purpose: in small samples they are more stable than fitted weights and keep interpretation straightforward [59]. These choices align the survey side with the log analogues in Chapter 4.

We show one association map using effect-size measures matched to variable types. Spearman's rank correlation (ρ) summarises ordered/numeric pairs; categorical pairs use Cramér's V on a $[0, 1]$ scale [6]. For ordered/numeric versus categorical (for example, usage_index by Q3 Plan type), we add the correlation ratio η to respect scale differences [2, 41]. Co-selection among tasks uses the Jaccard index, a standard overlap measure for

multi-label data [61]. Confidence intervals accompany the main displays; for proportions we use Wilson intervals [26].

We then reduce to a few dimensions using an exploratory factor/PCA workflow on usage_index, Q10 Usage timing (for trend checks), Q11 Prompt length (scored bands), Q12 Task breadth, family-specific subtask breadths (Q13–Q17), and Q18 Study/work share. Retention uses parallel analysis plus a scree check; an oblique rotation (e.g., OBLIMIN) allows related dimensions to correlate [13]. In practice we retain two to three axes that read naturally as (i) Usage intensity, (ii) Study/work orientation, and (iii) Portfolio breadth. We compute scores once and reuse them for Sdon so later contrasts refer to the same survey-defined constructs.

We relate factor scores to Q3 Plan type, Q4 Device, Q5 Status, Q6 Field and Q2 Age group using ordinary least squares (OLS) with HC3 robust standard errors. Ordered attitudes (Q19 Importance; Q20 Paid-only use, optionally collapsed to a binary in plots) are modelled with ordered logit or (for the binary collapse) logit; we present marginal changes in predicted probabilities rather than log-odds [55]. For non-normal contrasts on usage_index we also show a rank-sum test as a robustness check [16].

Alongside these models we report prevalence per Q12 Task family and subtask breadth within families (Q13–Q17), show which families are often chosen together (Jaccard co-selection), and connect breadth back to behaviour: Spearman's ρ relates task breadth to usage_index and to Q11 Prompt length bands; a rank-sum test compares usage_index between those who do versus do not select a given Q12 family. Q10 Usage timing is cross-tabulated with device, status and plan; Cramér's V summarises association strength and standardised residuals flag which cells drive any pattern [6]. This completes a compact, survey-only map of self-reported use. The same scores and definitions are then held fixed when section 5.2 compares Sdon and Llogs and when section 5.3 ranks and explains gaps.

5.2 SQ2: Survey–Log Convergence

SQ2 compares Sdon (n=24) and Llogs (n=24) on like-for-like indicators in the same units (weekly rates, dayparts, prompt-length bands, task codes). The composite usage_index is scaled on the pooled matched set (N=48) so both cohorts share one metric; this avoids re-scaling artefacts when distributions are shown side by side.

For numeric usage (Q7–Q9), means are fragile with small and skewed samples, so we use the Wilcoxon–Mann–Whitney rank-sum test. We summarise the size of any shift with the Hodges–Lehmann median difference in original units (sessions/week; minutes/session) and attach 95% bootstrap CIs. A rank effect (rank-biserial r or Cliff's δ) puts dominance on a $[-1, 1]$ scale. We also add two simple shape checks—the two-sample KS and a 1-Wasserstein distance—to see whether whole distributions differ, not just the middle [33, 46, 25, 16].

For Q10 Usage timing and Q11 Prompt length, we compare cohort shares per category with Wilson 95% intervals [26]. Differences are reported as $\Delta p = p_{\text{Sdon}} - p_{\text{Llogs}}$ with Newcombe score-type CIs, which remain stable at small n [29]. A single table-level number (Cramér's V) summarises overall association, and for Q11 we add a linear-by-linear trend across the ordered bands [6, 2]. The residual Varies too much to say is kept separate. For Q10 Usage timing we apply the dominance rule from Chapter 4: a daypart is assigned if

it captures at least one third of session starts; otherwise the label is Anytime throughout the day.

Tasks (Q12–Q17) are multi-label by design. For the six main families (Q12 Task families) we show prevalences with Wilson CIs and cohort differences with Newcombe CIs. We compare portfolio breadth—the number of Q12 families a person uses—using a rank-sum test and HL shift because breadth is a skewed count [16, 6]. For subtasks (Q13–Q17) we repeat the same within each family and gate logs on the parent family to keep like-for-like. Beyond single bars, we also check whether the overall mix aligns: Spearman's rank-correlation compares the ordering of families by prevalence, and a bounded Hellinger distance summarises how far the mixes are when many bars move a little [5].

We apply the same uncertainty and multiple-comparison rules as section 5.4. Effects are shown with 95% CIs, and we control false discovery at q=0.10 within three families: (i) Q7–Q11, (ii) Q12 options, and (iii) each subtask family (Q13–Q17) [7]. Plots mirror this plan: ECDF/violin views with HL shifts for Q7–Q9; stacked shares and Δp with intervals for Q10–Q11; per-option Δp (Q12 Task families) plus a compact breadth contrast; and a small panel with rank-correlation and Hellinger distance.

5.3 SQ3: Gaps and Subgroups

SQ3 ranks where Sdon and Llogs differ most and relates those components to survey subgroups. The aim is descriptive: show clear effect sizes in familiar units and then see which profiles sit higher or lower on those same components; no causal claims [52].

For numeric/ordinal usage (Q7–Q9 and the standardised usage_index) we report the HL median difference (Sdon − Llogs) with 95% bootstrap CIs, a rank effect (r or δ), and KS shape checks; where helpful we add a 1-Wasserstein distance [25, 34, 16, 33, 46]. For Q10–Q11 (categorical/ordinal), we compare cohort shares with Wilson intervals and Δp with score-type CIs, add Cramér's V for the table and a linear-by-linear trend for Q11, and keep Varies too much to say separate [26, 29, 54, 2]. For tasks, we mirror the survey at two levels: per-family prevalences and breadth (Q12), then gated subtasks (Q13–Q17), plus a pattern view with Spearman's ρ and Hellinger distance [5].

Inside the survey (N=93), we fit simple models with links matched to the outcome: OLS with HC3 for continuous indices and midpoint-coded usage (usage_index, Q7–Q9); ordered logit for ordered outcomes (Q11 bands; Q19 attitudes), reporting marginal effects; and logit for binaries (each Q12 family), reporting odds ratios and marginal effects in percentage points. Predictors are plan (Free/Plus/Pro), device (laptop/phone/mixed), status (student/working/both/other), field (STEM/business/H&SS/creative/other) and age band, with optional behavioural covariates (usage_index and Q12 task breadth) where useful. Coefficients are standardised where applicable, shown with 95% CIs, and interpreted as effect sizes, not pass/fail p-values [59]. These models explain who in the survey tends to sit higher or lower on the components that showed the largest Sdon–Llogs gaps.

Because Sdon and Llogs are independent and self-selected, we benchmark Sdon against the full Survey Sample on background variables (Q1–Q6; Q18–Q20): per-category Δp with Wilson intervals and Cramér's V as a compact strength-of-association measure [26, 54]. We also show that conclusions are stable under reasonable alternatives: (i)

different midpoints for open-ended bands (e.g., top-bin $\times 1.5$ for Q7–Q8), (ii) rare-category handling (collapsing ultra-rare subtasks or fields), and (iii) the episode-split sensitivity for log-derived durations/rates from Chapter 4. If an effect changes sign under a reasonable alternative, we flag it explicitly [28].

Results appear as (i) ranked gap bars with 95% CIs and BH–FDR marks; (ii) a compact table listing the impact measures used for ranking (HL in units, rank effect r/δ, V, and Δp); and (iii) coefficient/marginal-effects plots for the subgroup models.

5.4 Tests, uncertainty & reporting conventions

All tests are two-sided at α=0.05. We take an estimation-first approach: report effect sizes with 95% confidence intervals and treat p-values as diagnostics (shown, not used as pass/fail thresholds) [59]. Multiple comparisons are handled with Benjamini–Hochberg false-discovery-rate control at q=0.10 within three families: (i) Q7–Q11, (ii) Q12 options, and (iii) each subtask family (Q13–Q17) [7].

For numeric and ordinal usage (Q7–Q9; usage_index) we use Wilcoxon–Mann–Whitney and report HL median differences in natural units (sessions/week; minutes/session) and rank effects (rank-biserial r or Cliff's δ). We check distributional shape with KS and, where helpful, a 1-Wasserstein distance [16, 23, 6, 33, 46, 25].

For categorical outcomes we compare cohort shares rather than means. Category proportions use Wilson 95% intervals [26], and differences use Newcombe score-type intervals $\Delta p = p_{\text{Sdon}} - p_{\text{Llogs}}$ [29]. Overall association in contingency tables is summarised by Cramér's V on $[0, 1]$ [54]. For ordered bands (Q11) we add a linear-by-linear trend test; Varies too much to say is kept separate and, in a sensitivity run, omitted to check stability [2].

Uncertainty is handled with non-parametric bootstrap intervals (10,000 resamples; fixed seed) for HL and other statistics without small-sample closed forms; binomial share and Δp intervals follow the score-based formulas above. Analyses are available-case: we report the used N per item, do not impute missing values, and rely on the conservative bounds and proxy rules defined in Chapter 4 for the log side (for example, duration winsorisation) [27]. Units and coding are fixed: Q8 per day; Q9 in minutes per session; Q10 uses three dayparts on a Europe/Amsterdam time base with a dominance rule (a label is assigned if it captures $\geq 1/3$ of session starts; otherwise Anytime throughout the day); Q11 uses three prompt-length bands (sentence $\leq$20 words; short paragraph 21–60; multiple paragraphs $>$60) and assigns Varies too much to say if no band reaches $\geq 1/3$. Tasks (Q12–Q17) stay multi-label at the option level; breadth counts the number of families (Q12) and subtasks (within families) selected [61]. On the log side, aggregation thresholds are fixed: Q12 requires $n \geq 5$ and share $\geq 10\%$; Q13–Q17 require $n \geq 3$ and share $\geq 10\%$, and are gated on their Q12 parent.

6

Results

This chapter reports results aligned with the three sub-questions from Chapter 5. We start with the participation funnel and cohort balance, then map the self-reported landscape (SQ1), compare survey donors with logs donors on like-for-like indicators (SQ2), and identify the largest gaps and which survey subgroups align with higher or lower values (SQ3). We close with sensitivity and robustness checks. Throughout we emphasise effect sizes with confidence intervals and keep units consistent with Chapter 4.

6.1 Self-reported landscape (SQ1)

This section does two simple jobs. First, it shows which survey items move together so we can reuse one compact "intensity" score later in §6.2. Second, it checks whether differences mainly come from behaviour rather than who people are. We use the full survey only (N=93). Measures follow earlier definitions: usage items stay in their original bands and, where helpful, we also show midpoints (Q7 Sessions/week, Q8 Sessions/day, Q9 Session length). Q11 Prompt length stays ordered. Task breadth counts how many Q12 Task families a respondent selects. We first map associations, then reduce them to a few dimensions, and finally relate those dimensions to plan, device, status, field and age. Results are descriptive; full tables are in the appendix.

Association map

Figure 6.1 is a heatmap, each cell shows how strongly pairs of survey items move together on a common $[0, 1]$ scale (absolute Spearman's ρ for ordered/numeric pairs, Cramér's V for categorical–categorical, and η for ordered/numeric versus categorical) [61]. Darker cells mean a stronger link. Note: the map shows strength only (absolute values), not direction; it is not designed to display negative vs. positive signs. A rough guide for reading the colours: around 0.6 and above is strong; 0.3–0.6 is moderate; below 0.3 is weak [6].

Two things stand out. First, Q7–Q9 (how often and how long) move together and line up with Q11 Prompt length and task breadth. In plain terms, people who say they use ChatGPT more also tend to write longer prompts and report a wider set of task families; the links are strong rather than marginal.

Second, background variables (Q2 Age group, Q3 Plan type, Q4 Device, Q5 Status, Q6 Field) have smaller and mixed connections to behaviour. Co-selection among Q12 Task families matches intuition: Writing & communication pairs with Brainstorming/fun, and

Coding pairs with Language/translation, while Other co-selects weakly, as is typical for a residual category.Full cell estimates with intervals are in Table D.4; Q12 co-selection is in Table D.1.

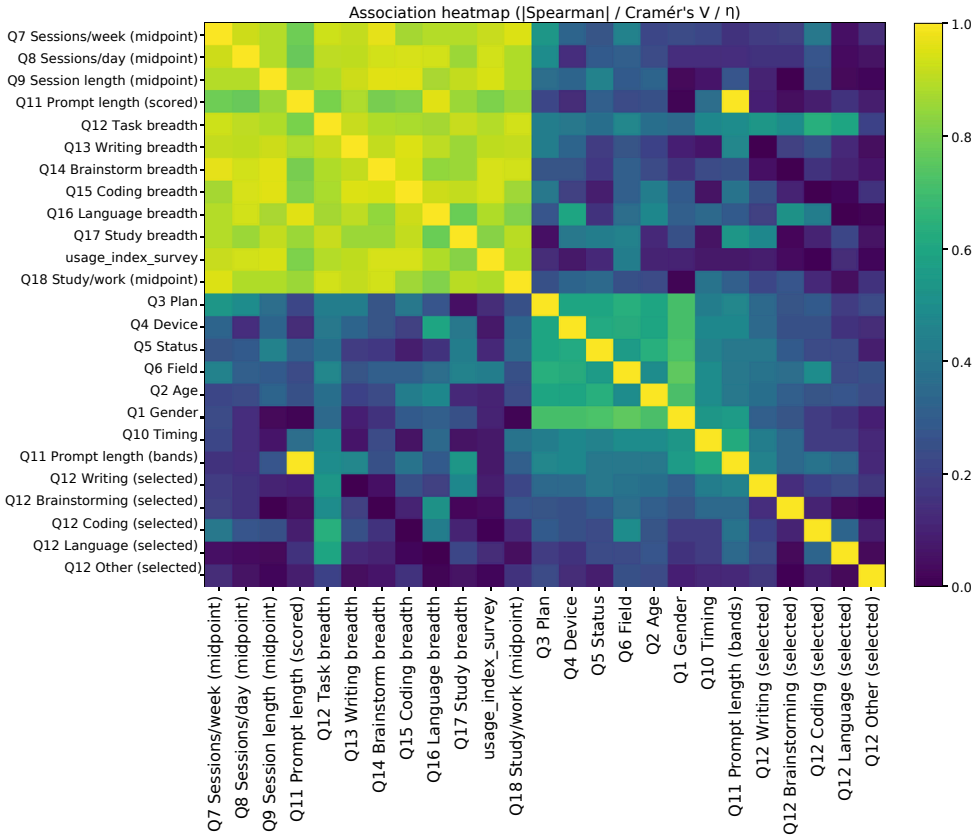

Figure 6.1: Association heatmap (S, N=93). Darker cells indicate stronger association on a common $[0, 1]$ scale.

Dimensions retained and what they mean

We want a small, stable set of axes we can reuse in §6.2–§6.3. Figure 6.2 shows a standard scree/parallel analysis: the first three observed eigenvalues sit above a random baseline; the fourth does not. We therefore retain three components using the oblique factor/PCA workflow [13]. In simple terms, they map to:

- **Usage intensity** — a "how much" axis: more sessions per week (Q7 Sessions/week), more per day (Q8 Sessions/day), and longer sessions (Q9 Session length). Q11 Prompt length shifts from one-liners towards short paragraphs as intensity rises. Moving up means using ChatGPT more often and for longer, and usually typing more than a single short line.
- **Study/work orientation** — where ChatGPT sits in daily routines: more study-leaning versus more work-leaning use, anchored by Q18 Study/work share (with Q10 Usage timing as a lighter cue). This is a tilt, not a volume knob.
- **Portfolio breadth** — how widely people range across task types: the number of Q12 Task families and subtask breadth (Q13–Q17). This is largely independent from intensity: someone can be broad but low-intensity, or narrow but heavy-intensity.

Rotated loadings and communalities are in Table D.3; removing the residual Q11 category ("Varies too much to say") does not change the meanings.

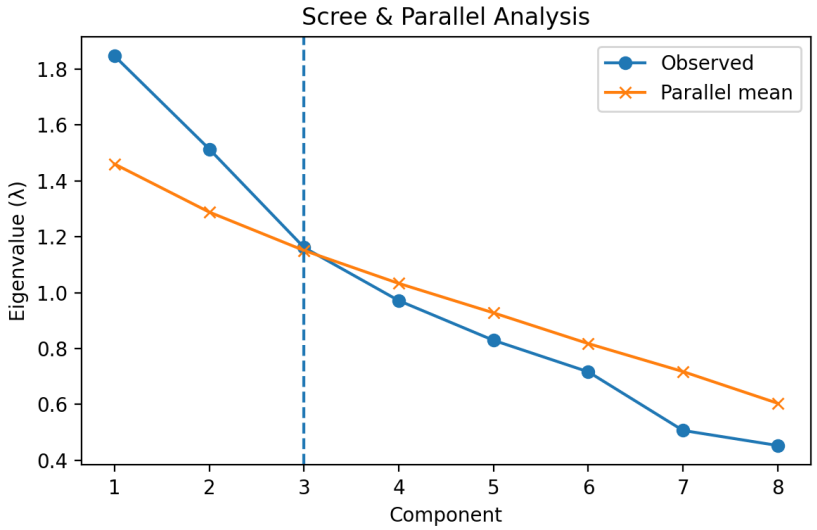

Figure 6.2: Scree and parallel analysis (S, N=93). We retain three components because their observed values sit above the random baseline.

Who scores higher or lower on these dimensions?

Next we relate each dimension to Q3 Plan type (Free/Plus/Pro), Q4 Device (laptop/-phone/mixed), Q5 Status, Q6 Field and Q2 Age group using OLS with HC3 standard errors [50]. Figure 6.3 displays the top coefficients with 95% CIs. Read this panel as a check: coefficients help us see whether any single profile dominates the dimensions. In practice, effects are modest—useful for interpretation, not for prediction. Full models appear in Table D.6–Table D.8.

Three plain takeaways:

1. **Intensity is only weakly tied to profiles:** The largest lift appears for Free plan; smartphone-only is slightly positive but uncertain. Paid plans (Plus/Pro) are not systematically higher in this convenience sample. Heavy users show up in every plan and device group.

2. **Orientation behaves like a tilt, not a volume knob:** Field and device shift where the axis points (for example, H&SS higher; laptop lower), without simply mirroring intensity.

3. **Breadth is independent from intensity:** Breadth is modestly lower among women and STEM; Plus users also show a small reduction. Trying many task families is not the same as using the tool a lot.

Attitudes line up with these patterns: higher intensity goes together with higher reported importance (Q19 Importance) and with a higher chance of continuing if access were paid-only (Q20 Paid-only use).

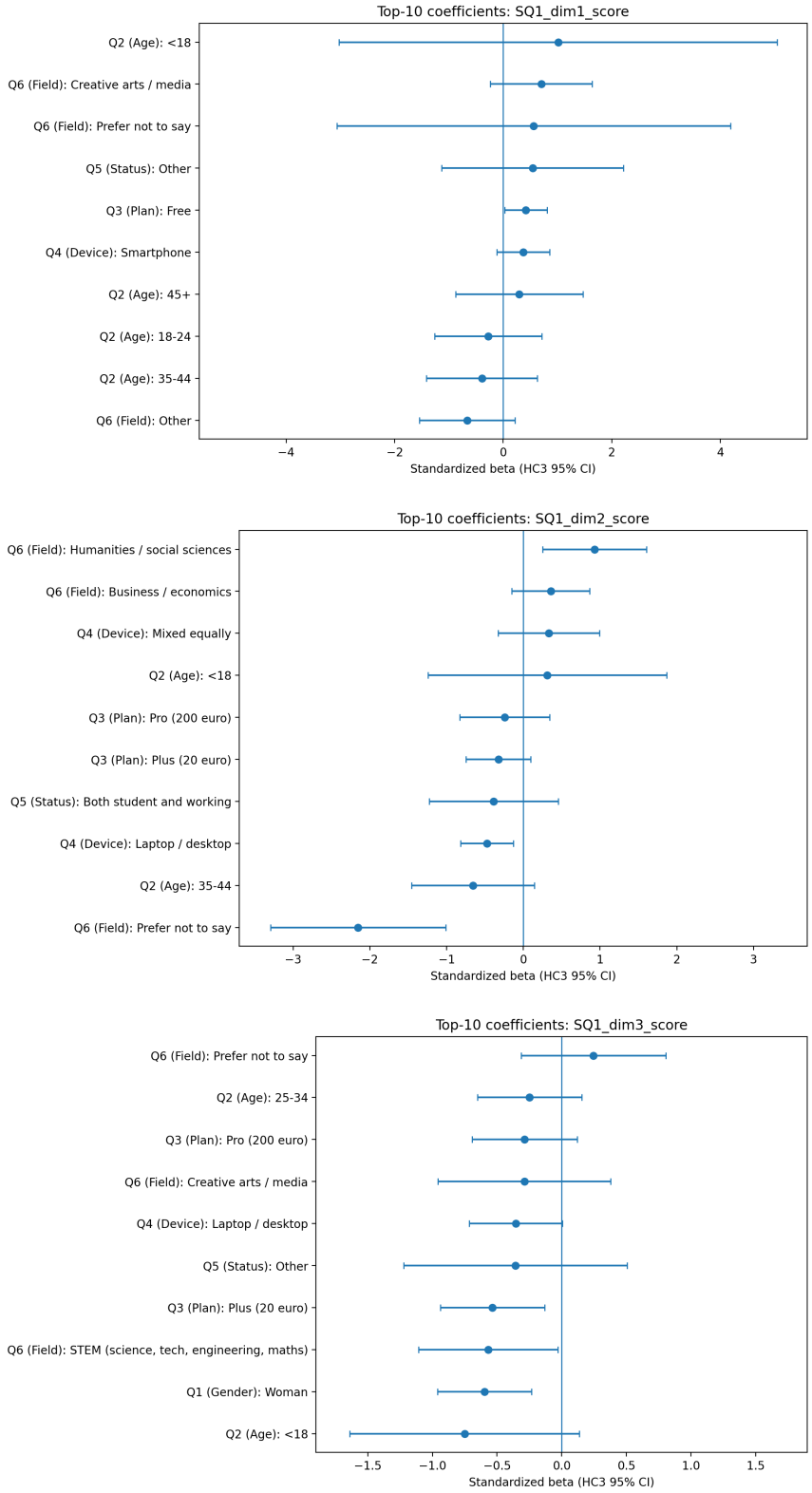

Figure 6.3: Profile associations for the three survey dimensions. Bars show estimates with 95% CIs; we display the top ten predictors per panel. Full models: Table D.6–Table D.8.

How to read these results

1. The survey provides a clear "how much" signal: Q7–Q9 and Q11 move together, so a single composite intensity score is meaningful.
2. Orientation and breadth are separate from intensity: someone can be study-leaning or work-leaning, and broad or narrow, at any usage level.
3. Profiles explain only a small part of the variation: differences by plan or device exist but are modest; they will not, on their own, explain survey–log gaps.

With that in place, section 6.2 puts survey donors next to logs donors on the same units and the same survey-defined constructs [61, 50, 6, 13].

6.2 Survey–log convergence (SQ2)

This section answers to the following question: do the survey-defined indicators line up with what we observe in logs when measured the same way? We place the survey-donor subset (Sdon, n=24) next to the independent logs donors (Llogs, n=24) on like-for-like frames (intensity, timing, prompt form, task portfolio). All contrasts are unpaired and distributional. For numeric outcomes we report Hodges–Lehmann (HL) shifts in natural units [25]; for categorical outcomes we show differences in shares (Δp in percentage points) with Newcombe score-type intervals [29]; and we control multiplicity with Benjamini–Hochberg FDR at q=0.10 [7]. Figure 6.4, Figure 6.5, Figure 6.6 and Figure 6.10 show the main displays; full tables appear in full tables are in the appendix (Table D.9–Table D.17). Unlike SQ3, which ranks the largest gaps and relates them to profiles, SQ2 is a side-by-side comparison of the distributions themselves.

Numeric usage

Figure 6.4 shows HL shifts (Sdon − Llogs) for Q7 Sessions/week, Q8 Sessions/day, Q9 Session length, and the composite usage index. Each dot is the estimated median difference in the original units; bars are 95% CIs. Values to the right mean Sdon sits higher than Llogs in that unit. We also check whether the shapes differ, not just the medians, using a two-sample KS and a 1–Wasserstein distance (Table D.9) [33, 46].

- **Q7 Sessions/week:** Survey donors report and show slightly more weekly sessions—on the order of a couple of extra sessions per week ($\approx +2.5$)—while the overall distribution looks very similar across cohorts.
- **Q8 Sessions/day:** Typical-day activity is essentially the same in both cohorts.
- **Q9 Session length:** Session lengths overlap broadly; medians line up.
- **Composite usage index:** Taken together, Sdon scores a little higher on overall intensity ($= +0.51$ pooled z), but the shift is modest.

Overall, numeric differences are small to moderate, and most intervals include small shifts (Table D.9). The larger contrasts show up elsewhere: in how people phrase prompts and in which task families they use.

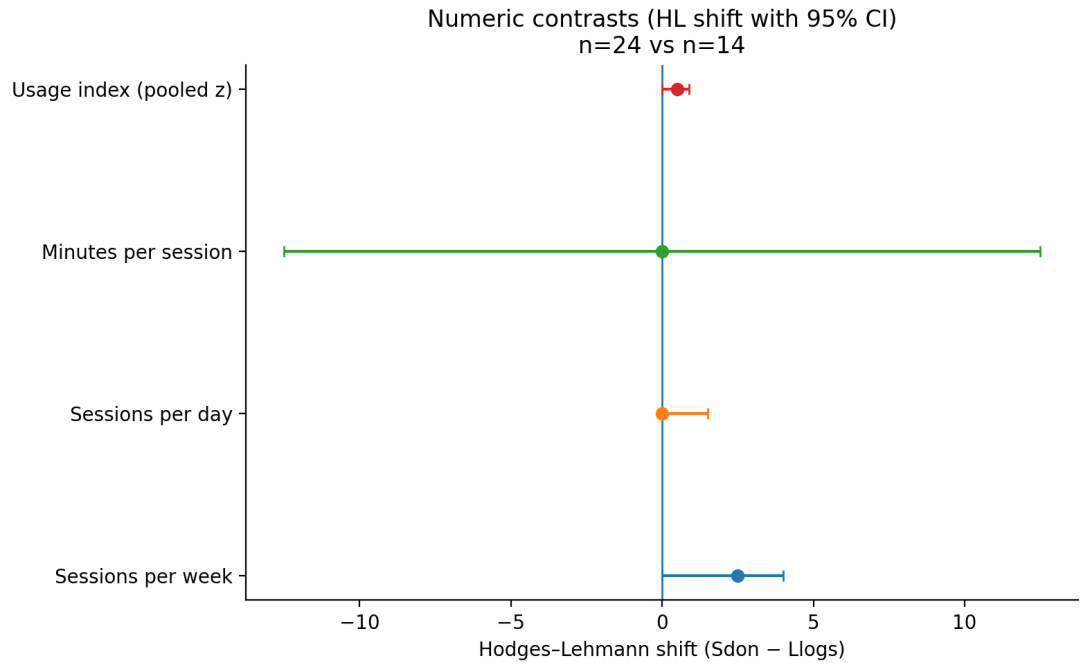

Figure 6.4: Numeric contrasts (Sdon − Llogs).

Q10 Usage timing

In Figure 6.5 the daypart labels are very similar across cohorts. Sdon shows a small tilt towards "Anytime throughout the day", mirrored by a small dip for "During work/study hours", and "Evenings" is essentially identical. Using the same dominance rule (at least a third of session starts) on both sides leads to the same story: timing is not a major source of difference.

Q11 Prompt length

Figure 6.6 shows large, one-direction differences. "A short paragraph" is much more common in Sdon (+54.2 percentage points), while "One short sentence" is much more common in Llogs (−41.7 points). "Multiple paragraphs" shows only a small lift for Sdon, and "Varies too much to say" is lower in Sdon. In plain terms, survey donors say they tend to write fuller, paragraph-level prompts; logs donors more often use one-liners.

Q12 Task families

Figure 6.7 summarises differences across the six task families and compares portfolio breadth. Two families drive most of the gap: Coding/programming and Language/translation are much less common in Sdon (each lower by roughly −37.5 and −41.7 percentage points). Writing & communication is a little higher, and Brainstorming/fun is about the same; Study/exam is very low in both cohorts. The residual "Other" shows a large negative gap by design (see Chapter 4). Portfolio breadth differs sharply: the median person uses 2 families in Sdon versus 4 in Llogs—a two-family gap that signals broader use on the logs side.

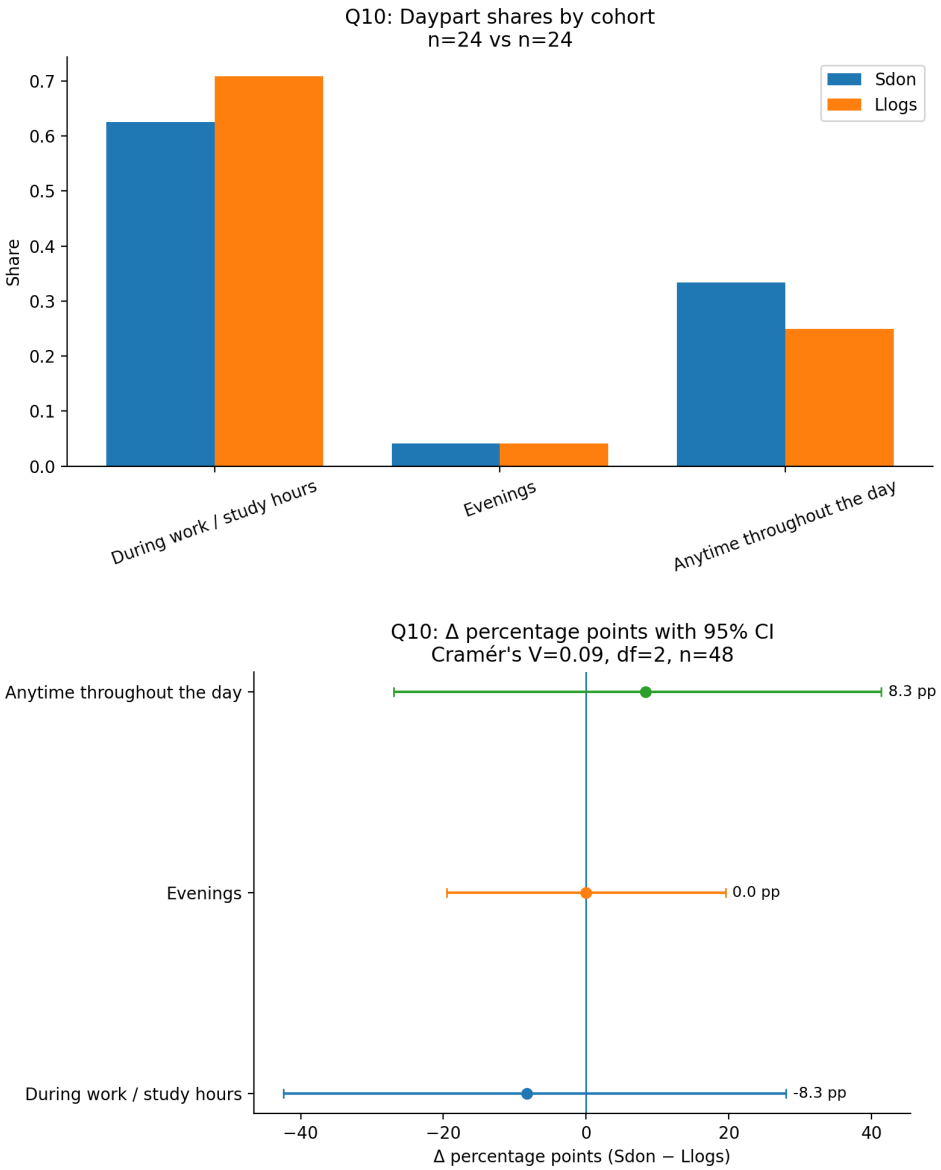

Figure 6.5: Q10 daypart shares (top) and Δp with 95% CIs (bottom). Table-level association V=0.09 (n=48).

Q13–Q17 Subtasks

Subtasks are computed only for donors who selected the parent family ("gated"), so comparisons are like-for-like. Two examples illustrate where the family-level gaps come from.

Q16 Language subtasks: Sdon is much lower on conversational practice/role-play and on vocabulary drills (the largest drops, on the order of −70 to −80 percentage points), and also lower on pronunciation. By contrast, Sdon is higher on improving grammar or style in a target language. In short, Sdon's language use leans towards polishing texts; Llogs leans towards interactive practice and drills (Figure 6.8).

Q15 Coding subtasks: Sdon is much higher on debugging existing code and on converting code between languages, and lower on explaining code/concepts and on writing unit tests. Here Sdon focuses on hands-on debugging and translation, while Llogs more often covers

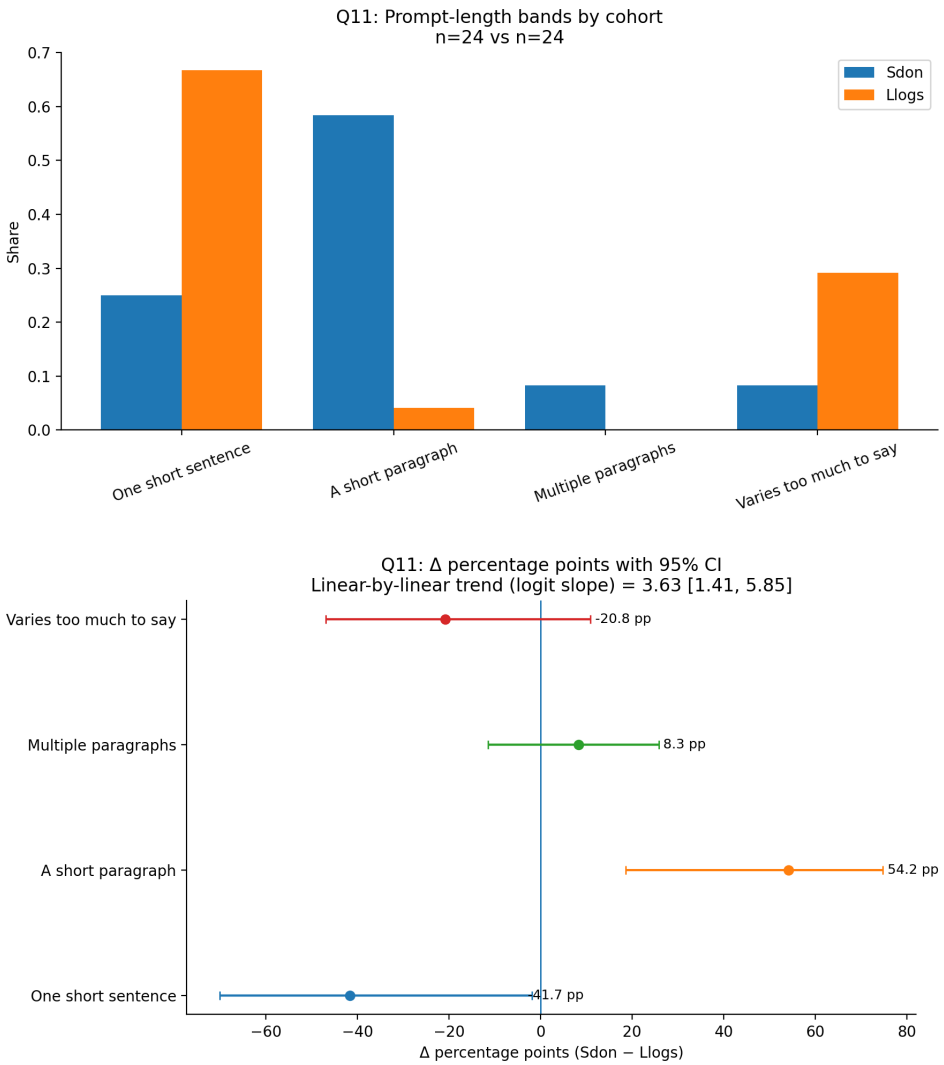

Figure 6.6: Q11 prompt-length shares (top) and Δp with 95% CIs (bottom). Bands are ordered; strong table-level association (V=0.65).

explanation and scaffolding (Figure 6.9). Appendix Figure E.1, Figure E.2 and Figure E.3 show similar tilts for Writing, Brainstorming/fun and Study/exam.

Pattern view

Figure 6.10 plots Sdon prevalence against Llogs prevalence for each Q12 family. Points do not sit on the diagonal: the rank ordering barely lines up and the overall mixes differ substantially. We show this view to stress that it is not just a few categories moving, but a broader reshaping of the portfolio.

How to read these results

Three takeaways.

1) Levels are broadly similar, while form and portfolio differ: Numeric intensity is only a little higher in Sdon and session length aligns; the big shifts are in prompt form (paragraphs versus one-liners) and in task breadth and mix.

2) Subtasks show the direction of travel: In language tasks Sdon emphasises grammar/style over conversational and drill-type uses; in coding Sdon leans into debugging and language

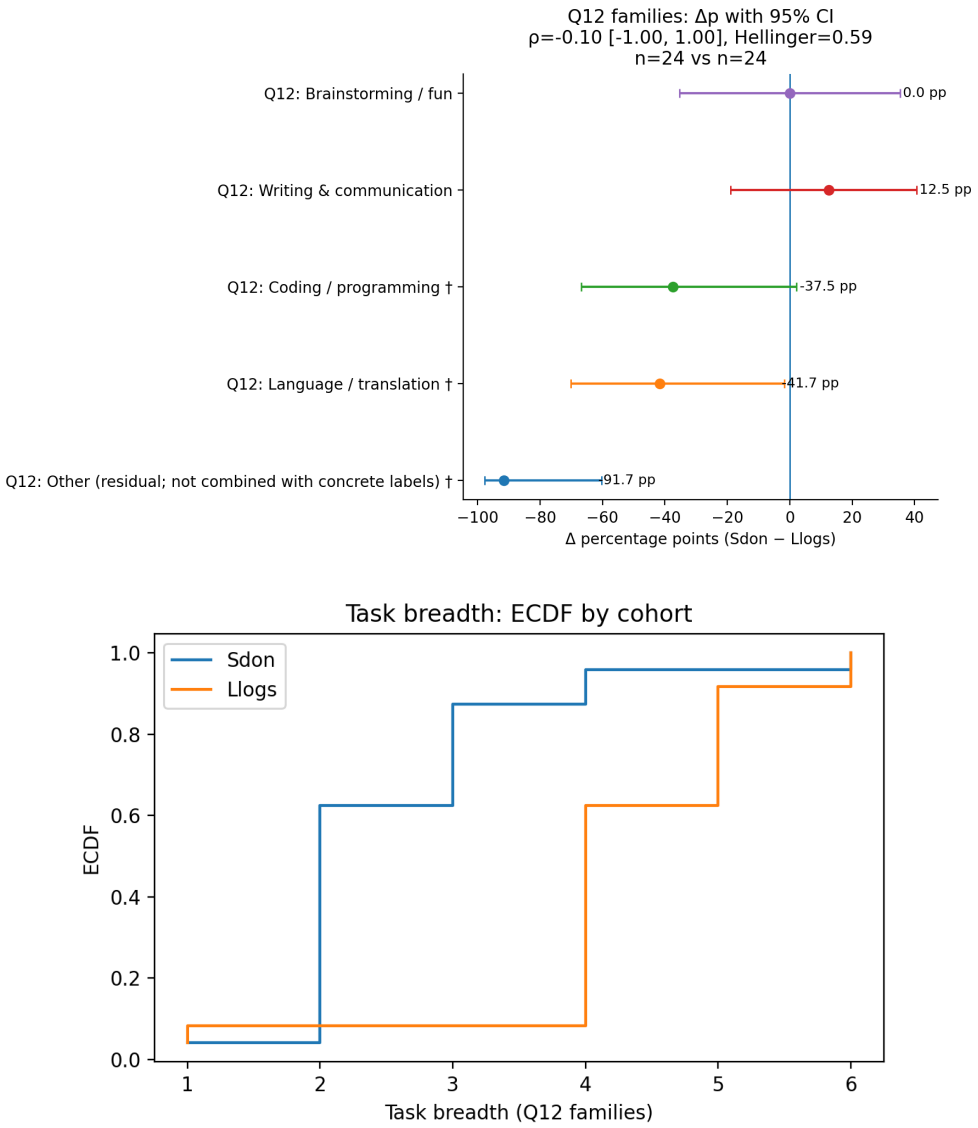

Figure 6.7: Q12 family differences.

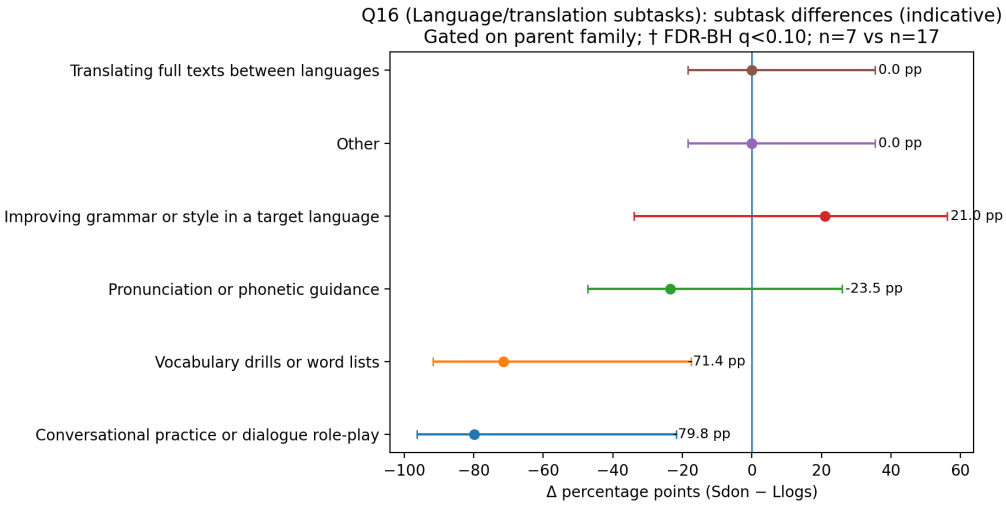

Figure 6.8: Q16 (Language/translation).

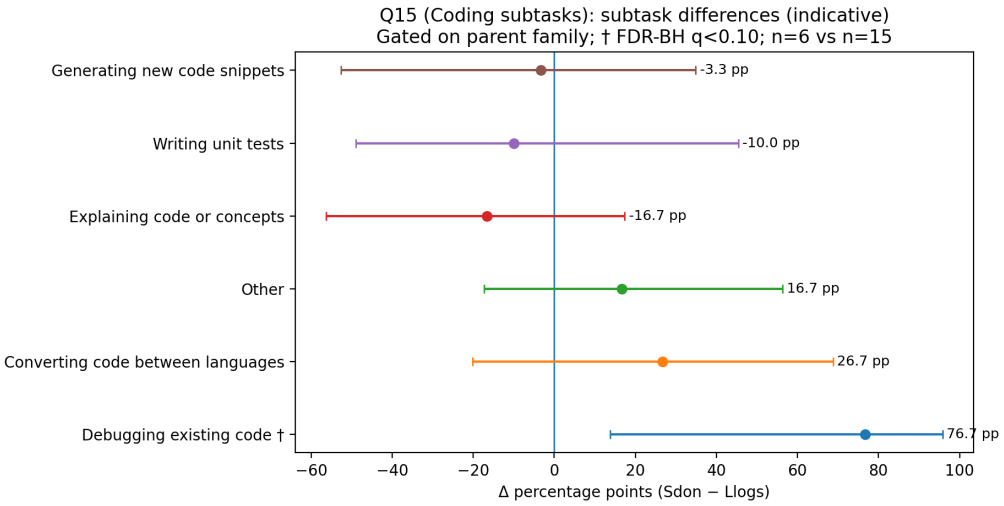

Figure 6.9: Q15 (Coding).

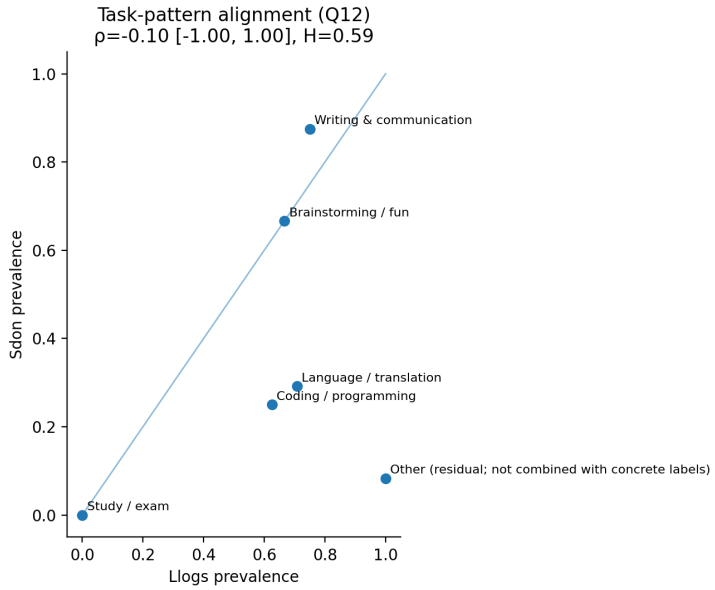

Figure 6.10: Task-pattern alignment (Q12). Spearman rank correlation and Hellinger distance relative to the diagonal reference.

conversion. Writing, brainstorming and study subtasks in the appendix show the same tilt away from short, interactive items.

3) Definitions matter: Timing labels and units match Chapter 4. “Other” is a residual category by design, and the large gap there is definitional, not substantive.

These contrasts identify where the cohorts differ most; section 6.3 ranks those gaps and asks which survey subgroups sit higher or lower on the same components.

6.3 Gaps and subgroups (SQ3)

This section ranks where Sdon and Llogs differ most and shows which survey subgroups tend to sit higher or lower on those same components. Effects are matched to the scale of each outcome (HL shifts for numeric, Δp in percentage points for categorical) and

presented with 95% CIs [25, 29]. We keep the frames from Chapter 4 and section 6.2 to ensure like-for-like comparisons, and control multiplicity with Benjamini–Hochberg within outcome families [7].

Top gaps

Figure 6.11 ranks the largest gaps: the top panel shows numeric HL shifts; the bottom panel shows Q12 Δp with 95% CIs. Category-level gaps for Q11 were shown in Figure 6.6.

The largest numeric gap is Q7 Sessions/week, where Sdon does a little more per week ($= +2.5$). The composite usage index is modestly higher for Sdon (HL $\approx +0.5$ pooled z). Q8 Sessions/day and Q9 Session length sit near zero. In short, everyday activity and session length look alike across cohorts; any intensity edge for Sdon sits mainly in weekly frequency and is small to moderate.

The biggest categorical differences appear in Q11 Prompt length (Figure 6.6): "A short paragraph" is about +54 points for Sdon, while "One short sentence" is about −42 points. In plain terms, survey donors gravitate to paragraph-level prompts, whereas logs donors more often use one-liners.

Two families dominate the Q12 ranking (Figure 6.11, bottom): Language/translation and Coding/programming are much less common in Sdon (each lower by roughly forty points). Writing & communication is slightly higher, and Brainstorming/fun is similar. "Other" shows a very large negative gap by definition (see Chapter 4). Portfolio breadth differs sharply as well: the median person uses 2 families in Sdon vs. 4 in Llogs (a two-family gap; see Table D.15). Beyond intensity, the cohorts diverge in what they do: the logs cohort spreads use across more families and is especially higher in coding and translation.

Where within families do gaps come from?

Subtasks clarify which activities drive the family-level differences (gated on the parent family; see section 6.2).

Q16 Language subtasks: Sdon is much lower on Conversational practice / dialogue role-play (−79.8 pp) and on Vocabulary drills or word lists (−71.4 pp), and also lower on Pronunciation or phonetic guidance (−23.5 pp). By contrast, Sdon is higher on Improving grammar or style in a target language (+21.0 pp). The survey-donor side leans towards text polishing, while the logs side leans towards interactive practice and drills (Figure 6.8).

Q15 Coding subtasks: Sdon is much higher on Debugging existing code (+76.7 pp) and on Converting code between languages (+26.7 pp), and lower on Explaining code or concepts (−16.7 pp) and Writing unit tests (−10.0 pp). The survey-donor side focuses on hands-on debugging and translation, whereas the logs side more often covers explanation and tests (Figure 6.9). Appendix Figure E.1–Figure E.3 (Writing, Brainstorming/fun, Study/exam) show similar tilts away from short, playful or drill-type items and towards summarising/explaining, consistent with the Q11 shift.

Who sits higher or lower on key gaps

We relate the main gap components to profiles using survey-only models with HC3 standard errors; coefficients are descriptive, not causal [50]. Read these models as a check on composition: if a single profile dominated a gap, it would show up here. Full tables are in Table D.19–Table D.22; Figure 6.12 summarises the coefficients for Q11 Prompt

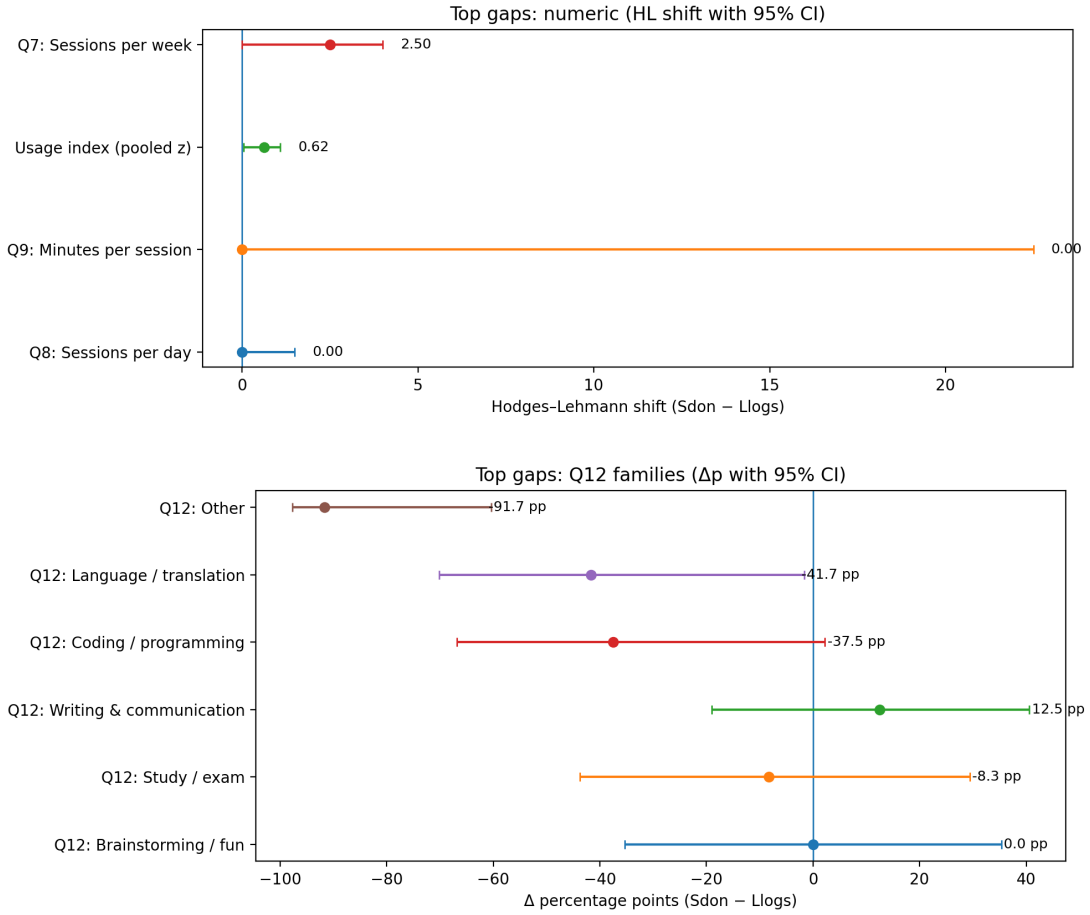

Figure 6.11: Top gaps by family: numeric HL shifts (top) and Q12 Δp with 95% CIs (bottom).

length and Q12 Task families along intensity.

For intensity, the strongest positives are Plus ($\beta \approx 0.52$) and Pro ($\beta \approx 0.28$), with a smaller positive for "both student and working" ($\beta \approx 0.24$). Field and age effects are modest. Heavier users show up across profiles, but paid plans are more likely to sit higher on the composite intensity.

For the probability of choosing "A short paragraph", smartphone use and the 18–24 age band are positive; "Other status" and "younger than 18" are negative. Longer prompts are thus more common among younger and phone-first respondents in this sample.

For Coding, field markers outside STEM (H&SS/Other/Business) are negative and Plus is positive; for Language/translation, Student and "Both student and working" are negative with a small positive for age 25–34. Selection into coding and translation therefore tracks field and role more than plan or device.

Representativeness of Sdon

Sdon is broadly similar to the full survey on background composition, with small to moderate differences on plan, device, status and field and larger differences on attitudes. This suggests composition differs somewhat but is unlikely to account for the large Q11/Q12 gaps on its own.

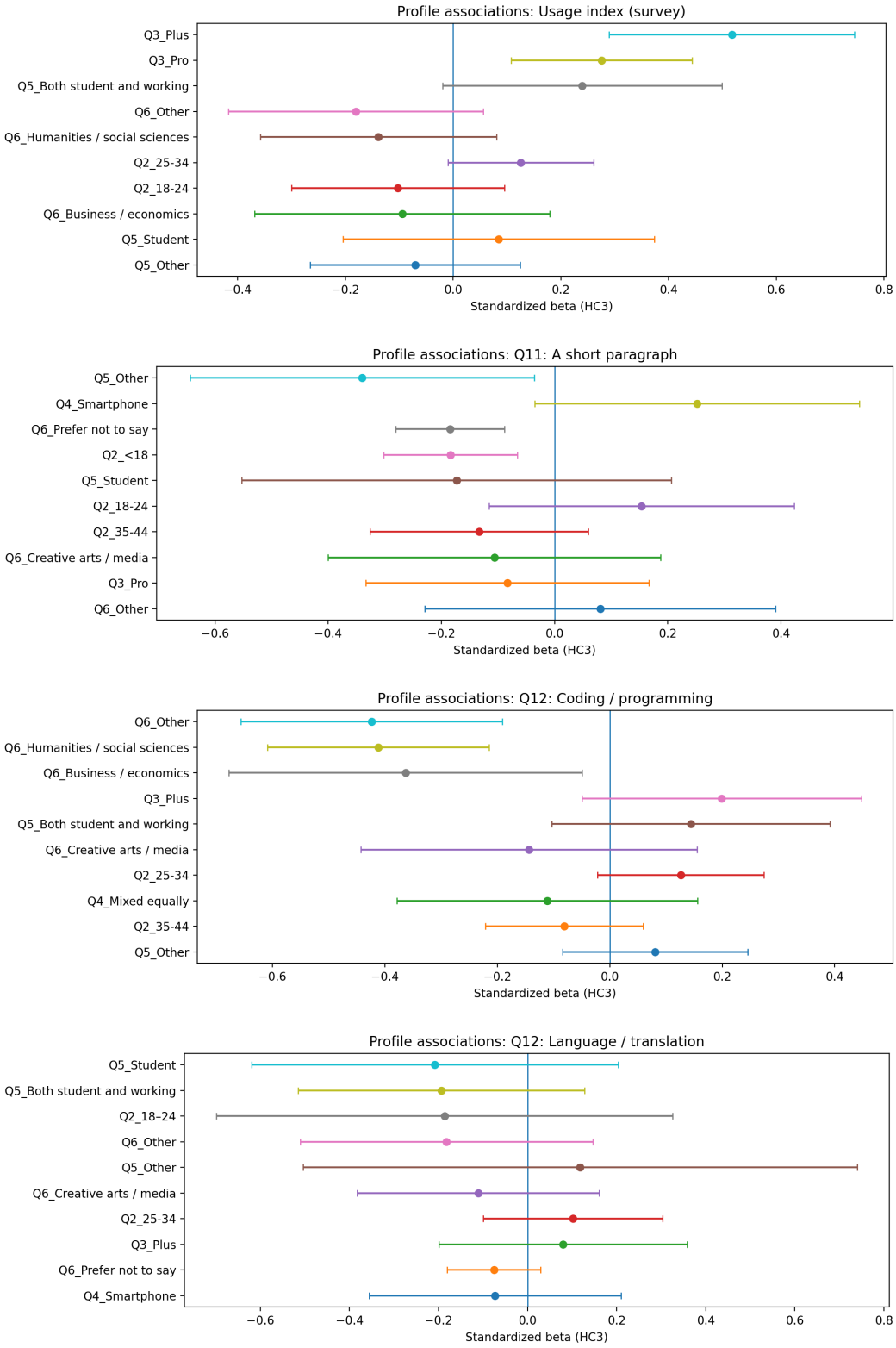

Figure 6.12: Profile associations for key gap components

Sensitivity and robustness

We check that conclusions do not hinge on specification choices. Wider top-bin midpoints do not change directions for Q7 Sessions/week (12→15), Q8 Sessions/day (6→7) or Q9 Session length (> 60: 75→90); see Table D.25. Available-case handling for Q9 Session length reduces n on the logs side and widens CIs but leaves medians aligned. All contrasts and models use the same units, Amsterdam time base and gating as in Chapter 4. Where relevant, shape diagnostics mirror section 6.2 (two-sample KS and 1–Wasserstein) [33, 46].

Key Takeaways

1) The biggest numeric gap is in weekly frequency; everyday activity per day and session length are alike.

2) The strongest categorical gaps sit in prompt form and task mix: Sdon shifts to paragraphs and a narrower set of families; Llogs spreads across more families and is higher in coding and translation.

3) Profiles help to read these differences (for example, paid plans and intensity; younger and phone-first for paragraphs), but they do not fully explain them. Mix and form remain the main points of divergence.

6.4 Summary and link back to the research question

Across SQ1–SQ3 the picture is consistent. SQ2 showed that levels of use align reasonably well across sources, while SQ3 organised where the biggest gaps lie and who tends to sit higher or lower on those gaps. On levels of use, the two cohorts look much alike; on how people use the tool, they pull apart.

Where the cohorts differ most is form and portfolio. Sdon more often uses paragraph-level prompts, while Llogs more often uses one-liners. Llogs also spreads activity across more task families (median 4 vs. 2) and is especially higher in Coding and Language/translation; Sdon is narrower and relatively more Writing. Inside families, the tilt is clear: for Language, Sdon focuses on grammar/style, while Llogs emphasises conversational practice, drills and pronunciation. For Coding, Sdon focuses on debugging and code conversion, while Llogs more often involves explanation and tests. Profile models help to read these patterns: higher intensity is more common among paid plans, longer prompts are more common among younger and phone-first respondents, and Coding/Language selection tracks field and role more than plan or device. These profiles do not fully explain the gaps.

Answering the RQ

How does ChatGPT usage in anonymised logs compare to what users report?

On how much people use the tool (frequency and duration), surveys and logs broadly agree. The biggest differences lie in how the tool is used. Self-reports describe longer, more elaborated inputs and a narrower set of task families; the logs show broader, more interactive and more technical use. These differences are not driven by timing or by one simple profile split, which points to a real gap between reported and observed practice.

Implication

Short, one-line and interactive or technical use is easy to miss in self-reports. Combining surveys with log-based indicators gives a fuller picture: the survey covers intensity well; the logs add detail on form and portfolio breadth. The concluding chapter places these findings in context, notes limitations, and turns them into practical guidance for instrument design and usage monitoring.

7

Discussion

This chapter interprets the main results, connects them to prior work on survey–trace alignment and everyday LLM use, and draws out what the patterns mean for measurement and practice. The guiding question stays practical: how does ChatGPT usage observed in anonymised logs compare to what users report in surveys, and which components align or diverge? We keep the stance descriptive: cohorts are independent by design, the sample is convenience-based and small, and effects are reported with uncertainty rather than as population estimates.

Two headline patterns organise the discussion. First, levels of use are broadly similar across cohorts: typical-day activity and minutes per session overlap, and the only clear numeric difference is a modest shift in weekly frequency towards the survey-donor group. Second, the largest gaps appear in how people interact (prompt form) and what they use the tool for (task breadth and mix). These differences are not explained by timing labels and are only weakly related to composition, which points to real gaps between reported and observed practice rather than artefacts of who donated logs.

7.1 Discussion of findings

What aligned: a workable signal for "how much"

Across numeric indicators, survey donors and logs donors look much alike. Typical-day sessions and session length line up; the median shift sits close to zero on both measures. Weekly frequency is somewhat higher among survey donors (Hodges–Lehmann $\approx +2.5$ sessions/week; 95% CI $[0.0, 4.0]$), but shapes remain comparable. Read plainly: when framed in broad, interpretable units, respondents recall how much they use ChatGPT with acceptable fidelity. This matches validation work on digital behaviour showing that coarse frequency/duration bands travel better from memory than fine-grained counts [32, 60].

Where cohorts pulled apart

The strongest categorical differences are in prompt form. Survey donors most often describe a short-paragraph input; logs show a much higher share of one-liners. Concretely, "short paragraph" is over-represented on the survey side ($\Delta p \approx +54\,\text{pp}$), while "one short sentence" dominates more on the logs side ($\Delta p \approx -42\,\text{pp}$); the table-level association is strong (Cramér's $V \approx 0.65$; see Figure 6.6). This signals distinct interaction styles: survey

narratives gravitate to fuller, worked-out prompts, whereas the logs capture many terse nudges and checks.

Task breadth and mix diverge as well. The median survey donor names two main families; the median logs donor uses four (see Figure 6.7). Family-level gaps concentrate in Coding/programming and Language/translation, both substantially more prevalent in logs; Writing & communication is slightly more common on the survey side, while Brainstorming/fun is similar. In short, logs reveal a broader and more technical or interactive portfolio, not just a uniform rescaling of the same tasks.

Within-family subtask tilts make this texture concrete. In Language/translation, survey donors gravitate to polishing (grammar/style), while logs donors more often show practice and drills (conversational role-play, vocabulary, pronunciation; Figure 6.8). In Coding, survey donors emphasise debugging and code conversion; logs donors show more explanation and scaffolding (Figure 6.9). Read together with the prompt-length split, these tilts indicate that the logs contain many short, iterative exchanges that are easy to under-weight when someone is asked for their "usual" use. Family-level inferences are supported by validation (micro-F1 ≈ 0.63; Figure 4.2 and Table 4.3); subtask inferences are directional rather than precise (micro-F1 ≈ 0.45–0.54), and we treat them accordingly.

Why these gaps are plausible

Three mechanisms, all consistent with prior measurement work, credibly account for the form/portfolio gaps:

(1) Recall and salience. When asked for a typical prompt or usual uses, respondents compress varied behaviour to a memorable prototype. Longer, production-like prompts are easier to recall and justify; one-line nudges and micro-checks fade from memory. Surveys therefore over-select paragraph-level inputs and under-select technical or interactive micro-uses; logs count every exchange across the full horizon [32, 60].

(2) Blended workflows. Real sessions often chain tasks (for example, outline → translate → refine), and "language help" or "code explanation" may not be perceived as distinct tasks when answering a checklist. A multi-label coding of prompts captures these overlaps explicitly in logs; single-pass survey recall flattens them. The broader log portfolios and the consistent within-family tilts are exactly what a blended, opportunistic workflow would yield [61].

(3) "Usually" versus "observed across the horizon". Survey wording ("usually") encourages naming a few salient families. On the log side, a family only counts for a donor after clearing minimum support (for example, $n \geq 5$, share $\geq 10\%$), which biases against breadth inflation. That breadth is still larger in logs strengthens the interpretation that surveys under-enumerate secondary but regular families.

What did not drive the gaps

Two potential alternative explanations do not fit the data.

(1) Timing. Timing labels are essentially the same across cohorts (weak table-level association; "Evenings" identical), which rules out a simple "when" story (see Figure 6.5). If the cohorts use the tool at similar times but show different prompt forms and portfolios, the difference lies in interaction style, not schedule.

(2) Composition. Background differences between the survey-donor subset and the full survey are small to moderate (see Table D.23). Profile models (plan, device, field, age) help read who sits higher or lower on some components (for example, paid plans and intensity), but they do not erase the prompt-form or breadth gaps. That is, who people are explains little; how they describe use versus what logs observe remains the main source of divergence.

7.1.1 How this study compares to prior work

Our main alignment is with validation studies that compare self-reports to digital traces. Meta-analyses and recent smartphone work find that broad frequency/duration measures tend to match logs more closely than fine-grained counts, while form and short, iterative interactions are often under-reported in surveys [32, 60, 45]. In our data, this is exactly what we see: Q7 Sessions/week, Q8 Sessions/day and Q9 Session length are broadly similar across sources (a small weekly lift for survey donors), whereas Q11 Prompt length and Q12 Task families show the largest gaps (paragraphs vs. one-liners; narrower vs. broader portfolios).

On adoption and typical activities, our survey taxonomy mirrors what broad student snapshots and early campus studies report. Student surveys emphasise writing/communication, language help and study support as recurring clusters [1, 4, 18, 24]. Large general polls point in the same direction for adults and teens: usage rose steadily between 2023 and 2025 [9, 10]. In that context, our Q12 Task families and the observed mix (more Writing on the survey side; more Coding and Language/translation in logs) fit the wider picture: surveys foreground longer, production-like prompts and a few salient families; logs surface many short, technical or interactive tasks.

For the workplace, landscape studies highlight rapid uptake with uneven patterns across roles and sectors [12, 30, 19]. Our side-by-side view is consistent with that texture: intensity (how much) aligns reasonably well, but form and portfolio (how and what) vary—especially Coding and Language/translation—suggesting that brief, task-specific uses are easy to miss in self-reports.

Methodologically, our donation-based, unpaired comparison follows recommended practice: keep survey and trace cohorts independent, align frames so each survey item has a like-for-like log analogue, and avoid one-to-one linkage for privacy [31, 47, 49, 14]. Relative to many telemetry-only studies, we add profiles (plan, device, status, field) from the survey; relative to survey-only studies, we add prompt-form and portfolio breadth from logs. Read plainly: our study sits between the two traditions—surveys for "who" and context; logs for "how" and "how widely".

Limitations also match prior work. Donation cohorts are selective and small, so effects are descriptive; subtask labels are informative about direction but less precise. Even so, the main alignment (levels) and the main divergences (form and mix at Q11/Q12) replicate the pattern others have documented in adjacent settings [32, 60].

7.2 Implications

This section translates the patterns above into concrete guidance for how to measure everyday LLM use and what organisations should monitor or support. The advice follows

directly from two stable findings: (i) levels of use align reasonably well between survey and logs; (ii) the biggest gaps live in input form and portfolio breadth or mix. We keep the focus practical and avoid restating results already shown.

7.2.1 Implications for measurement and method

Treat "use" as four separate facets—intensity, timing, input form and task portfolio—and design survey items so each has a like-for-like log analogue. In this study, typical-day sessions and minutes per session align closely across sources, and weekly frequency differs only modestly (Hodges–Lehmann shift $\approx +2.5$ sessions/week). Timing labels are also similar when broad dayparts share one local time base. These patterns justify keeping coarse, interpretable bands for frequency or duration and using the same bands when deriving log indicators. In practice: pre-specify weekly or daily frames, a local time base with a simple dominance rule for dayparts, and minute bands for sessions; mirror those frames in the export. This preserves interpretability and makes distributional contrasts straightforward.

The largest survey–log differences are categorical: survey donors gravitate to paragraph-level inputs while logs show many more one-liners (Cramér's $V \approx 0.65$; Figure 6.6); log donors also span more task families (median 4 vs. 2; Figure 6.7) with higher prevalence of Coding and Language/translation. Instruments that only ask "how often?" will miss these contrasts. In practice: include a prompt-length item with simple word-count bands that the log side can reproduce; and a compact checklist of main task families scored as multi-label, with portfolio breadth reported alongside prevalence.

"Usually" prompts respondents to name a few salient categories and longer, worked-out inputs; logs surface short, iterative prompts that get under-reported. Two small wording tweaks help: (i) anchor to a concrete horizon ("In the last 30 days, which of these have you done at least five times?") to align with log thresholds; (ii) allow a "Varies" option but keep it distinct so it does not dilute ordered bands. Our log aggregation counted a family for a donor only if it cleared $n \geq 5$ and $\geq 10\%$ of prompts—a conservative choice that still yielded broader portfolios in logs, underscoring the need to counter salience in surveys.

For free-text logs, we mapped prompts to the same task families respondents saw using a small, fixed codebook and a hybrid router (embedding-based prototype matching with a minimalist judge at boundaries). Family-level validation supports prevalence contrasts (micro-F1 ≈ 0.63; Table 4.3); subtask labels are informative about direction but less precise (micro-F1 ≈ 0.45–0.54). The implication is twofold: (i) keep the family space small and public (prototypes in an appendix) so others can audit and reuse; (ii) treat subtask differences as indicative unless supports are large. Publish thresholds and ambiguity rules so replication is possible.

Small samples and skewed usage benefit from robust estimators (medians for skewed numeric quantities; Wilson or Newcombe intervals for binomial shares; HL shifts for location). Thresholds used for timing labels (one-third dominance), inactivity splits (30 minutes) and portfolio inclusion (share and count minima) are design choices; state them up front and keep them identical across sources. Sensitivity checks can then verify that qualitative conclusions do not hinge on a single setting.

Comparative analysis worked here without one-to-one linkage: we sized a survey-donor

subset to the logs cohort and contrasted full distributions. This preserves anonymity while still revealing where sources agree (levels, timing) and where they diverge (form, portfolio). If linkage is not strictly necessary for the research aim, prefer unpaired and distributional designs.

7.2.2 Implications for organisations

Dashboards that only track "active users" or "sessions" risk a false sense of alignment: our study shows intensity aligns across sources, but the mode of interaction and the breadth of use do not. If you monitor LLM use, add two shape indicators: (i) the distribution of prompt-length bands and (ii) portfolio breadth (how many task families a person typically uses). Expect a heavier mass of one-liners and a broader task mix in logs than you will elicit from surveys; calibrate "healthy use" baselines accordingly.

Logs reveal many micro-tasks—quick translations, code explanations, vocabulary checks, terse summaries—and more technical or interactive families than respondents list. Training that only models long, polished prompts under-serves real behaviour. Teach short iterative prompting, chaining (for example, search → draft → revise), and code or translation "nudge" patterns alongside longer drafting. Inside families, emphasise the tilts we observe (for example, language practice and drills; code explanation and tests) so materials match everyday texture rather than idealised cases.

Because timing labels looked alike across cohorts on broad dayparts, "when" people use the tool is a weak lever relative to "how" they use it. Plan enablement and support independent of time-of-day variation; invest instead in shaping input form and portfolio breadth where the biggest gaps—and opportunities—sit.

If you need richer telemetry, consider user-centric donation with platform-native exports, tiered consent, data minimisation and no one-to-one linkage. Our funnel shows feasibility (93 surveys; 51 reached the donation page; 24 uploaded logs; Figure 3.1), with conversion depending on clear instructions and trust; analyses then proceed as unpaired, distributional comparisons. This model balances utility and governance and is suitable for pilots in education or the workplace.

When reporting usage to stakeholders, be explicit about what each indicator can and cannot say. Family-level prevalence contrasts are well supported; subtask splits are directional unless supports are large. Keep "Other" as a residual by design (not a substantive category), and present effect sizes with intervals rather than single-point claims.

7.3 Limitations and directions for future work

The design choices in this study were made to enable a clean, privacy-preserving comparison between self-reports and logs. Those choices also define the limits of what we can claim. We group these limits into four themes and pair each with concrete, feasible next steps.

(1) External validity and recruitment

Our evidence comes from a convenience sample recruited via an intranet post and Instagram (S: N=93; Sdon and Llogs: n=24 each). The pool is male-leaning and younger (about two-thirds aged 18–35), which reflects the reach of the channels (Figure 3.1; section 3.1).

All contrasts are therefore descriptive rather than population estimates, and subgroup models are read as associations, not causal effects (Chapters 3, 5, 6).

Future work. Broaden recruitment beyond networks to reduce skews and improve transportability: mix university lists, professional associations, and opt-in panels with pre-set quotas by age, role, and field; publish funnel metrics and composition alongside outcomes. Where feasible, run parallel cohorts in different settings (for example, specific departments or business units) and non-Western contexts to test whether the form or portfolio gaps replicate. When quotas are impossible, use transparent post-stratification against neutral benchmarks strictly for descriptive re-weighting, not for model-based inference.

(2) Privacy-preserving, unpaired design

Re-identification risk was reduced by design: via platform-native exports, data minimisation and no one-to-one linkage. The trade-off is that we cannot decompose within-person survey–log differences (all comparisons are unpaired, distributional). This constraint is deliberate and central to donor trust, but it limits what we can say about individual recall-bias mechanisms or the stability of self-reports over time (Chapters 3–4).

Future work. Explore tightly governed, opt-in linkage variants that preserve the spirit of this study: (i) a double-consent, short-lived join token administered inside a secure enclave so raw identifiers never leave donor control; (ii) ephemeral, local joins that produce only pre-agreed aggregates (for example, donor-level bands) before keys are destroyed; and (iii) third-party safe-room audits in which an independent steward verifies the join and releases only cell-count–protected outputs. Any such design would require additional ethics review and explicit safeguards against the known risks of re-identification (hashing $\neq$ anonymity), but would enable within-person convergence tests that are out of scope here.

(3) Measurement frames and task coding

Frames were aligned ex-ante (weekly or daily rates; dayparts; prompt-length bands; multi-label task taxonomy) and thresholds were set for interpretability (for example, Q12 family counted at $n \geq 5$ and $\geq 10\%$; subtasks at $n \geq 3$ and $\geq 10\%$; dominance rules for timing and prompt-length). These are sensible design choices, but they are still choices. Family-level labels validate well (micro-F1 ≈ 0.63), while subtask labels are noisier (micro-F1 ≈ 0.45–0.54), so subtask contrasts are read as directional, not precise (see Chapter 4, Table 4.3; see also the per-label panels in the appendices). The Other category is a residual by design and exclusive on the log side, which affects its apparent cohort difference (Chapter 4).

Future work. Two paths can sharpen fidelity without sacrificing transparency. First, richer form features: add conversational texture (turn count per session, back-and-forth depth), scaffolding markers (explain $\rightarrow$ try $\rightarrow$ revise sequences), and light tool-use flags to complement word-count bands, so "one-liner" versus "paragraph" carries interaction context. Second, codebook and threshold refinement: expand prototypes with active learning (human review of uncertain prompts); replace hard minima with Bayesian shrinkage at the donor level for rare but recurrent subtasks; and report label-wise calibration plots alongside F1. Both improvements keep the pipeline auditable while reducing boundary effects that can under- or over-state breadth.

(4) Small n and temporal scope

The matched cohorts are small (24 vs. 24), and the donated horizons vary. We therefore emphasised effect sizes with intervals, rank-based contrasts, and FDR control; several numeric gaps centre near zero with wide CIs (for example, minutes/session), while categorical gaps (prompt form; portfolio breadth) are large and consistent (HL shift for Q7 Sessions/week $\approx +2.5$ sessions/week; Cramér's $V \approx 0.65$ for Q11 Prompt length) (Chapter 6). Cross-sectional timing also means we cannot separate seasonal or policy effects from idiosyncratic behaviour.

Future work. Field a panel donation with two to four waves per donor to track within-person stability of intensity, form and breadth; this allows random-effects models that separate person-level differences from week-to-week variation. Pre-register a short list of primary endpoints (for example, Q11 band share; Q12 breadth) to control multiplicity upfront and retain the estimation-first stance. Where feasible, embed light natural experiments (policy changes, training roll-outs) and measure coupled outcomes (perceived productivity, quality, creativity) on the same frames to test how form and breadth travel with perceived value in real tasks rather than constrained benchmarks.

Summary

The limits are tractable: broaden and diversify recruitment to improve transportability; keep privacy by design while prototyping consented, enclave-based linkage for within-person checks; enrich form features and calibrate the task pipeline to tighten prevalence estimates; and add a panel structure to separate stable patterns from short-run noise. Each step extends the present study without discarding the core commitments that made it feasible: like-for-like frames, multi-label realism, and privacy-first governance.

8

Conclusion

This thesis compared what people say they do with ChatGPT to what anonymised platform exports show, using a donation-based design without linking individuals. Because the cohorts are small and independent, we compare groups and report differences with uncertainty rather than make population-wide claims.

SQ1. What dimensions structure self-reported use (intensity, study/work orientation, breadth), and how do they relate to plans, devices, roles and fields?

The survey data show a clear usage core. Q7 Sessions/week, Q8 Sessions/day and Q9 Session length move together and also line up with Q11 Prompt length and the number of Q12 Task families people select. On top of that core sit three readable components: Intensity (how often/how long, with a shift from one-liners towards short paragraphs), Study/Work orientation (a tilt rather than a volume knob), and Portfolio breadth (how many task families a person ranges across). Crucially, breadth is not just intensity by another name. Background factors (plan, device, field, age) play a smaller role—heavy and light users appear in every profile—so later survey–log gaps are unlikely to be explained by composition alone.

SQ2. To what extent do survey constructs correspond to log indicators on like-for-like frames (rates, dayparts, prompt-length bands, task families)?

On "how much", the two sources broadly agree. Weekly frequency is somewhat higher among survey donors (about +2.5 sessions per week), while typical-day activity and minutes per session look alike. Timing labels are also similar, so "when" people use the tool is not the main difference. The clearest gap is *form*: survey donors report many more short-paragraph prompts (about +54 percentage points), whereas the logs show many more one-liners (about −42 points). Portfolios diverge too: logs span more task families (median 4 vs. 2), especially Coding/programming and Language/translation. In short, the survey captures "how much"; the logs add detail on "how" and "how widely".

SQ3. Which components differ most between survey donors and logs donors, and which survey subgroups sit higher or lower on those components?

Ranked by size, the biggest gaps are categorical: prompt form (paragraphs vs. one-liners), prevalence of Coding and Language/translation, and overall portfolio breadth. Numeric gaps are smaller (a modest weekly-frequency shift; other level metrics near zero). In survey-only models, intensity is higher among paid plans; longer prompts are more common

among younger and phone-first respondents; and selecting Coding or Language/translation tracks field/role more than plan or device. These patterns help read the gaps but do not explain them away.

Based on the findings of the three sub-questions, this thesis aimed to answer the following overarching RQ:

How does ChatGPT usage in anonymised logs compare to what users report?

When both sources use the same frames, they largely agree on how much people use ChatGPT, apart from a small weekly edge for survey donors. They diverge on how and what: logs show many short, one-line, interactive or technical exchanges and a broader mix of tasks; surveys emphasise paragraph-level prompts and a narrower portfolio. These differences are not about timing or one simple subgroup—they reflect a real gap between what people remember and what the logs record.

Implications and closing

For measurement, treat usage as multi-facet—intensity, timing, form and portfolio—and, with small samples, report ranges rather than single numbers. Logs should complement, not replace, surveys: donation-based exports sharpen form and breadth; surveys add profiles and context. For governance and MOT practice, a privacy-by-design setup (platform-native export, minimal data, no record linkage) makes responsible monitoring feasible. Our convenience sample means results are descriptive, but using both lenses together gives a practical way to see not just how much people use LLMs, but also how and for what.

9

Bibliography

[1] A. A. Abdulla and et al. “Higher education students’ perceptions of ChatGPT: A global study of early reactions”. In: *Frontiers in Education* 9 (2024). URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11798494/.

[2] Alan Agresti. “Measures of Nominal–Ordinal Association”. In: *Journal of the American Statistical Association* 76 (1981), pp. 524–529. DOI: 10.1080/01621459.1981.10477679. URL: https://users.stat.ufl.edu/~aa/articles/agresti_1981.pdf.

[3] T. Alashoor et al. “Privacy Concerns and Data Donations: Do Societal Benefits Matter?” In: *MIS Quarterly* 49.2 (2025), pp. 429–464. URL: https://kambizsaffari.com/research/.

[4] C. Baek, T. Tate, and M. Warschauer. “ChatGPT seems too good to be true: College students’ use and perceptions of generative AI”. In: *Computers and Education: Artificial Intelligence* 7.3 (2024), p. 100294. URL: https://escholarship.org/uc/item/2gf4p9hh.

[5] Bayesia S.A.S. *Hellinger Distance — Key Concepts (BayesiaLab)*. Definition and properties for discrete prevalence vectors. 2023. URL: https://www.bayesia.com/bayesialab/key-concepts/hellinger-distance (visited on 08/30/2025).

[6] Mattan S. Ben-Shachar et al. “Phi, Fei, Fo, Fum: Effect Sizes for Categorical Data That Use the Chi-Squared Statistic”. In: *Mathematics* 11.9 (2023), p. 1982. DOI: 10.3390/math11091982. URL: https://www.mdpi.com/2227-7390/11/9/1982 (visited on 08/22/2025).

[7] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995). Author-hosted PDF of the original JRSS B article, pp. 289–300. URL: https://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini%20and%20Y%20FDR.pdf.

[8] Mohamad Azhari Bujang and Norsa’adah Baharum. “Guidelines of the minimum sample size requirements for Cohen’s Kappa”. In: *Epidemiology, Biostatistics and Public Health* 14.2 (2017). Guidelines on N for κ; formula-based planning. URL: https://www.researchgate.net/publication/320148141_Guidelines_of_the_minimum_sample_size_requirements_for_Cohen’s_Kappa (visited on 08/30/2025).

[9] Pew Research Center. *34% of U.S. adults have used ChatGPT, about double the share in 2023*. https://www.pewresearch.org/short-reads/2025/06/25/34-of-

us-adults-have-used-chatgpt-about-double-the-share-in-2023/. Accessed: 2025-08-06. June 2025.

[10] Pew Research Center. *About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023*. https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/. Accessed: 2025-08-06. Jan. 2025.

[11] Lingjiao Chen, Matei Zaharia, and James Zou. "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance". In: *arXiv preprint* (2023). DOI: 10.48550/arXiv.2305.05176. eprint: 2305.05176. URL: https://doi.org/10.48550/arXiv.2305.05176.

[12] Michael Chui et al. *The State of AI in 2023: Generative AI's Break-out Year*. McKinsey Global Survey report. McKinsey & Company. Aug. 1, 2023. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

[13] Anna B. Costello and Jason W. Osborne. "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis". In: *Practical Assessment, Research, and Evaluation* 10.7 (2005). Parallel analysis; scree; prefer oblique rotation. URL: https://openpublishing.library.umass.edu/pare/article/1650/galley/1601/view/ (visited on 08/30/2025).

[14] CSMaP NYU. *Beyond Red-Teaming: Facilitating User-Based Data Donation to Study Generative AI*. Policy brief. 2024. URL: https://csmapnyu.org/impact/policy/beyond-red-teaming-facilitating-user-based-data-donation-to-study-generative-ai.

[15] Elizabeth Mirekuwaa Darko, Manal Kleib, and Joanne Olson. "Social Media Use for Research Participant Recruitment: Integrative Literature Review". In: *Journal of Medical Internet Research* 24.8 (2022), e38015. DOI: 10.2196/38015. URL: https://www.jmir.org/2022/8/e38015.

[16] Michael P. Fay and Michael A. Proschan. "Wilcoxon–Mann–Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules". In: *Statistical Surveys* 4 (2010), pp. 1–39. DOI: 10.1214/09-SS051. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2857732/ (visited on 08/22/2025).

[17] Federal Trade Commission (FTC). *No, hashing still doesn't make your data anonymous*. FTC Tech Blog guidance on re-identification risks of hashing. July 2024. URL: https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/no-hashing-still-doesnt-make-your-data-anonymous (visited on 08/30/2025).

[18] J. von Garrel and J. Mayer. "Artificial Intelligence in studies—use of ChatGPT and AI-based tools among students in Germany". In: *Humanities and Social Sciences Communications* 10 (2023), p. 799. URL: https://www.nature.com/articles/s41599-023-02304-7.

[19] Anders Humlum and Emil Vestergaard. *The Unequal Adoption of ChatGPT Exacerbates Inequalities Among Workers*. Tech. rep. Working Paper 2024-50. Becker Friedman Institute, University of Chicago, 2024. URL: https://bfi.uchicago.edu/wp-content/uploads/2024/04/BFI_WP_2024-50.pdf.

[20] Information Commissioner's Office (ICO). *Introduction to anonymisation*. Guidance. 2021. URL: https://ico.org.uk/media2/migrated/2619862/anonymisation-intro-and-first-chapter.pdf.

[21] S. Mo Jones-Jang et al. "Good News! Communication Findings May Be Underestimated: Comparing Effect Sizes with Self-Reported and Logged Smartphone Use

Data". In: *Journal of Computer-Mediated Communication* 25.5 (2020), pp. 346–363. DOI: 10.1093/jcmc/zmaa009. URL: https://academic.oup.com/jcmc/article/25/5/346/5896236.

[22] Jae Hyun Lee et al. "Quantifying Smartphone Use: Objective Measures, Diurnal Patterns, and Validity against Self-Reports". In: *PLOS ONE* 12.1 (2017). Open access, e0169983. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5355019/.

[23] Katherine Meissel and Eno S. Yao. "Using Cliff's delta as a non-parametric effect size measure: An accessible web app and R tutorial". In: *Practical Assessment, Research, and Evaluation* 29.2 (2024). URL: https://openpublishing.library.umass.edu/pare/article/1977/galley/1980/view/.

[24] J. Nam. *56% of College Students Have Used AI on Assignments or Exams*. https://www.bestcolleges.com/research/most-college-students-have-used-ai-survey/. Accessed: 2025-08-06. Nov. 2023.

[25] National Institute of Standards and Technology (NIST). *Difference of Hodges–Lehmann*. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/diffhdle.htm. Dataplot Reference Manual (U.S. NIST). 2003. (Visited on 08/22/2025).

[26] National Institute of Standards and Technology (NIST). *7.2.4.1. Confidence intervals (binomial): Wilson and related methods*. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm. NIST/SEMATECH e-Handbook of Statistical Methods. 2012. (Visited on 08/22/2025).

[27] National Institute of Standards and Technology (NIST). *Engineering Statistics Handbook: Outliers (Exploratory Data Analysis)*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm. 2012.

[28] NCBI Bookshelf. *Sensitivity Analysis — Encyclopedia of Biostatistics*. Reference article, accessed 2025-08-18. URL: https://www.ncbi.nlm.nih.gov/books/NBK126178/.

[29] Robert G. Newcombe. "Interval estimation for the difference between independent proportions: Comparison of eleven methods". In: *Statistics in Medicine* 17.8 (1998). Recommends unpooled Wilson (Method 10) for Δp CIs, pp. 873–890. DOI: 10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I. URL: https://www.researchgate.net/publication/13687790_Interval_estimation_for_the_difference_between_independent_proportions_Comparison_of_eleven_methods (visited on 08/30/2025).

[30] OECD. *Using AI in the Workplace: Opportunities, Risks and Policy Implications*. Tech. rep. OECD, 2024. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/using-ai-in-the-workplace_02d6890a/73d417f9-en.pdf.

[31] Jakob Ohme and Theo Araujo. "Digital Data Donations: A Quest for Best Practices". In: *Patterns* 3.4 (2022), p. 100467. DOI: 10.1016/j.patter.2022.100467. URL: https://www.cell.com/patterns/fulltext/S2666-3899(22)00052-6.

[32] Douglas A. Parry et al. "A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use". In: *Nature Human Behaviour* 5.11 (2021), pp. 1535–1547. DOI: 10.1038/s41562-021-01117-5. URL: https://purehost.bath.ac.uk/ws/files/219658181/Final_Manuscript.pdf.

[33] John A. Peacock. *The two-dimensional Kolmogorov–Smirnov test*. Background on KS; distributional comparison via sup-norm. 2011. URL: https://www.researchgate.

net/publication/49399948_The_two-dimensional_Kolmogorov-Smirnov_test (visited on 08/30/2025).

[34] Pennsylvania State University, Eberly College of Science. *STAT 200: Lesson 4.3 — Introduction to Bootstrapping*. https://online.stat.psu.edu/stat200/lesson/4/4.3. Open online course material. n.d. (Visited on 08/22/2025).

[35] Pew Research Center. *Social Media Fact Sheet*. Age distribution of Instagram use; neutral demographic benchmark. 2024. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ (visited on 08/30/2025).

[36] N. Pfiffner and T. N. Friemel. "Leveraging Data Donations for Communication Research: Exploring Drivers Behind the Willingness to Donate". In: *Communication Methods and Measures* 17.3 (2023), pp. 227–249. URL: https://www.zora.uzh.ch/id/eprint/255389/9/Leveraging_Data_Donations_for_Communication_Research__Exploring_Drivers_Behind_the_Willingness_to_Donate.pdf.

[37] Nico Pfiffner et al. "Comparing Online Recruitment Strategies for Data Donation Studies". In: *Computational Communication Research* 6.2 (2024), pp. 1–19. DOI: 10.5117/CCR2024.2.8.PFIF. URL: https://journal.computationalcommunication.org/article/view/8602.

[38] Joelle Pineau et al. "Improving Reproducibility in Machine Learning Research: A Report from the NeurIPS 2019 Reproducibility Program". In: *Journal of Machine Learning Research* 22 (2021), pp. 1–20. URL: https://www.jmlr.org/papers/volume22/20-1344/20-1344.pdf.

[39] Afsaneh Razi et al. "Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection". In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA '22)*. ACM, 2022. DOI: 10.1145/3491101.3503569. URL: https://stirlab.org/wp-content/uploads/jan422-chi22extendedabstracts-22-1.pdf.

[40] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.pdf.

[41] Nicolas Robette. *Tutorial (English): Descriptive statistics — correlation ratio η^2*. GDAtools vignette; definition and interpretation of η, η^2. 2022. URL: https://nicolas-robette.github.io/GDAtools/articles/english/Tutorial_descr.html (visited on 08/30/2025).

[42] S. Rohr et al. *Non-Probability Surveys*. GESIS Survey Guidelines. Guideline recommending descriptive framing for non-probability samples. GESIS – Leibniz Institute for the Social Sciences, 2024. URL: https://www.gesis.org/fileadmin/admin/Dateikatalog/pdf/guidelines/non_probability_surveys_rohr_felderer_silber_daikeler_rossmann_schroeder_2024.pdf (visited on 08/30/2025).

[43] Takaya Saito and Marc Rehmsmeier. "The Precision–Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In: *PLOS ONE* 10.3 (2015), e0118432. DOI: 10.1371/journal.pone.0118432. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432.

[44] Salesforce. *More than Half of Generative AI Adopters Use Unapproved Tools at Work – Generative AI Snapshot: The Promises and Pitfalls of AI at Work*. https:

//www.salesforce.com/news/stories/ai-at-work-research/. Accessed: 2025-08-06. Nov. 2023.

[45] Michael Scharkow. "The Accuracy of Self-Reported Internet Use: A Validation Study Using Client Log Data". In: *Communication Methods and Measures* 10.1 (2016), pp. 13–27. DOI: 10.1080/19312458.2015.1118446. URL: https://www.tandfonline.com/doi/full/10.1080/19312458.2015.1118446.

[46] SciPy Developers. `scipy.stats.wasserstein_distance` *(documentation)*. 1-Wasserstein (Earth Mover's) distance; API reference. 2025. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html (visited on 08/30/2025).

[47] Henning Silber et al. "Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 185.S2 (2022), S387–S409. DOI: 10.1111/rssa.12944. URL: https://academic.oup.com/jrsssa/article/185/Supplement_2/S387/7069509.

[48] Philipp Singer et al. "Evidence of Online Performance Deterioration in User Sessions on Reddit". In: *PLOS ONE* 11.8 (2016), e0161636. DOI: 10.1371/journal.pone.0161636. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161636 (visited on 08/22/2025).

[49] A. Skatova and J. Goulding. "Psychology of personal data donation". In: *PLOS ONE* 14.11 (2019), e0224240. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6867598/.

[50] StataCorp LLC. *Heteroskedasticity-robust standard errors: some practical considerations*. Accessible overview; complements Long & Ervin (2000). Oct. 2022. URL: https://blog.stata.com/2022/10/06/heteroskedasticity-robust-standard-errors-some-practical-considerations/ (visited on 08/30/2025).

[51] M. Stureborg et al. "Crowdsourcing Hierarchical Multi-Label Annotations". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Shows parent–child gating improves efficiency and F1. 2023. DOI: 10.1145/3544548.3581431. URL: https://dl.acm.org/doi/10.1145/3544548.3581431 (visited on 08/30/2025).

[52] Xin Sun et al. "Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses". In: *BMJ* 340 (2010), p. c117. DOI: 10.1136/bmj.c117. URL: https://www.bmj.com/content/340/bmj.c117.

[53] Feliks Prasepta Sejahtera Surbakti. "Systematic Literature Review on Generative AI: Ethical Challenges and Opportunities". In: *International Journal of Advanced Computer Science and Applications* 16.5 (2025), pp. 307–315. URL: https://thesai.org/Downloads/Volume16No5/Paper_30-Systematic_Literature_Review_on_Generative_AI.pdf.

[54] UCLA Institute for Digital Research and Education (IDRE). *PROC FREQ: Annotated output (including Cramér's V)*. https://stats.oarc.ucla.edu/sas/output/proc-freq/. UCLA Statistical Consulting Group. n.d. (Visited on 08/22/2025).

[55] UCLA Statistical Consulting Group. *FAQ: How do I interpret the coefficients in an ordinal logistic regression?* Accessed 23 August 2025. n.d. URL: https://stats.oarc.ucla.edu/other/mult-pkg/faq/ologit/.

[56] UK Government. *Linking with anonymised data: How not to make a hash of it*. Guidance document. 2021. URL: https://www.gov.uk/government/publications/

joined-up-data-in-government-the-future-of-data-linking-methods/linking-with-anonymised-data-how-not-to-make-a-hash-of-it.

[57] United Nations Economic Commission for Europe (UNECE). *Guidance on Data Editing and Imputation (SDE 2022 Session 4 materials)*. Conference of European Statisticians. 2022. URL: https://unece.org/sites/default/files/2022-10/SDE2022_S4_Spain_Barrag%C3%A1n%20%26%20Salgado_D.pdf.

[58] R. Wang et al. *The Role of Privacy Guarantees in Voluntary Donation of Private Data for Altruistic Goals*. arXiv preprint arXiv:2407.03451. Accessed: 2025-08-06. 2024. URL: https://arxiv.org/html/2407.03451v1.

[59] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. "Moving to a World Beyond "p < 0.05"". In: *The American Statistician* 73.sup1 (2019), pp. 1–19. DOI: 10.1080/00031305.2019.1583913. URL: https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913 (visited on 08/22/2025).

[60] Alexander Wenz, Florian Keusch, and Rebecca L. Bach. "Measuring Smartphone Use: Survey Versus Digital Behavioral Data". In: *Social Science Computer Review* 42.5 (2024). Advance online publication pagination in some databases, pp. 1–20. DOI: 10.1177/08944393231159836. URL: https://journals.sagepub.com/doi/10.1177/08944393231159836.

[61] Min-Ling Zhang and Zhi-Hua Zhou. "A Review on Multi-Label Learning Algorithms". In: *IEEE Transactions on Knowledge and Data Engineering* 26.8 (2014), pp. 1819–1837. DOI: 10.1109/TKDE.2013.39. URL: https://palm.seu.edu.cn/zhangml/files/TKDE'13.pdf.

A

Consent and Ethics Materials

Survey Consent

You are invited to participate in a research study by Quinten Voncken, a Master's student at TU Delft, supervised by Dr. Savvas Zannettou. This study investigates how accurately users self-report their use of ChatGPT. Participation involves completing an anonymous online survey, taking approximately 2–5 minutes, about your usage frequency, tasks performed and attitudes toward ChatGPT. Your responses are anonymous, will be securely stored and your anonymous answers, with the answers of other respondents will be made publicly available at the end of the study. Participation is voluntary, and you can withdraw at any point without consequences. Proceeding with this survey indicates your consent to participate under these conditions. For questions, contact Quinten Voncken or Dr. Savvas Zannettou.

Thank you and Data Donation Consent

Thank you for completing the survey. You now have the opportunity to voluntarily donate your anonymised ChatGPT usage logs for further analysis and, if you choose to do so, enter a lottery to win a €50 Bol.com gift card. Participation in this step is completely optional and independent of your survey responses.

All submitted logs will be treated as personal data and handled in accordance with the European GDPR. Before analysis, logs will be processed to ensure full anonymization. We will not analyze or retain any personally identifiable information contained in the logs. Only general usage characteristics such as frequency, types of prompts and interaction patterns will be examined. The file is stored under a random code only, so neither we nor your school / employer can trace it back to you or to your survey responses. The anonymised data will be used solely for research purposes, securely stored, and accessed only by the research team at TU Delft. Your email address and raw log file will be permanently deleted at the end of the study (October 2025).

By submitting your anonymised logs, you explicitly consent to their use in this research under the conditions described. If you choose to participate and follow the instructions on the next page to upload your logs, you will automatically be entered into the lottery. The winner of the €50 Bol.com voucher will be randomly selected and notified via email on 23 July 2025.

If you have any questions or need more information, please contact Quinten Voncken or Dr. Savvas Zannettou.

B

Survey Instrument and Codebook

Part	Question	Answers
Part 1	Q1 - What is your gender?	Woman
		Man
	Q2 - Which age group do you belong to?	< 18
		18–24
		25–34
		35–44
		45+
	Q3 - Which ChatGPT plan are you currently on?	Free
		Plus (20 euro)
		Pro (200 euro)
	Q4 - Which device do you use most often to access ChatGPT?	Laptop / desktop
		Smartphone
		Mixed equally
	Q5 - What best describes your current status?	Student
		Working
		Both student and working
		Other
	Q6 - What best describes your main field of study or work?	STEM (science, tech, engineering, maths)
		Business / economics
		Humanities / social sciences
		Creative arts / media
		Prefer not to say
		Other

Part	Question	Answers
Part 2	⋆ Q7 - How many separate ChatGPT sessions did you have in the last 7 days?	0
		1–2
		3–5
		6–10
		More than 10
	⋆ Q8 - On a typical day, how many ChatGPT sessions do you start?	0
		1
		2–3
		4–5
		6 or more
	⋆ Q9 - On average, how long does a single ChatGPT session last?	Less than 5 minutes
		5–15 minutes
		15–30 minutes
		30–60 minutes
		More than 60 minutes
	⋆ Q10 - When do you most often use ChatGPT?	During work / study hours
		Evenings
		Anytime throughout the day
	⋆ Q11 - How long are your typical prompts?	One short sentence ($\leq$ 20 words)
		A short paragraph (21–60 words)
		Multiple paragraphs ($>$ 60 words)
		Varies too much to say
Part 3	⋆ ◇ Q12 - What tasks do you usually use ChatGPT? (multiple answers possible)	Writing & professional communication
		Brainstorming & personal ideas / fun
		Coding / programming help
		Language practice or translation
		Study revision / exam prep
		Other
	⋆ ◇ Q13 - If you chose “Writing & professional communication”, which sub-tasks do you use ChatGPT for?	Outlining ideas or slides
		Drafting full text

Part	Question	Answers
		Proof-reading / tone adjustment
		Summarising sources or meeting notes
		Adjusting style for different audiences
		I did not choose “Writing & professional communication”
		Other
	⋆ ⋄ Q14 - If you chose “Brainstorming & personal ideas / fun”, what kinds of prompts do you ask ChatGPT for?	Academic or research topics
		Business or marketing concepts
		Creative role-play, jokes, stories
		Hypothetical “what-if” scenarios
		Recommendations (books, movies, music)
		Trivia & general knowledge
		I did not choose “Brainstorming & personal ideas / fun”
		Other
	⋆ ⋄ Q15 - If you chose “Coding / programming help”, what coding sub-tasks do you use ChatGPT for?	Generating new code snippets
		Debugging existing code
		Explaining code / concepts
		Converting code between languages
		Writing unit tests
		I did not choose “Coding / programming help”
		Other
	⋆ ⋄ Q16 - If you chose “Language practice or translation”, what do you mainly use ChatGPT for?	Translating full texts between languages
		Improving grammar or style in a target language
		Vocabulary drills or word lists
		Conversational practice / dialogue role-play
		Pronunciation or phonetic guidance
		I did not choose “Language practice or translation”
		Other

Part	Question	Answers
	⋆ ◇ Q17 - If you chose "Study revision / exam prep", which study tasks do you use ChatGPT for?	Summarising lecture notes or readings
		Generating practice questions or quizzes
		Explaining difficult concepts in simple terms
		Reviewing flashcards / key terms
		I did not choose "Study revision / exam prep"
		Other
	Q18 - In the last month, what share of your total ChatGPT sessions were for study or work tasks?	0%
		1–25%
		26–50%
		51–75%
		76–100%
	Q19 - How important is ChatGPT for completing your study or work tasks?	Extremely important
		Somewhat important
		Neutral
		Somewhat not important
		Not important at all
	Q20 - If ChatGPT became paid-only tomorrow, would you still use it?	Yes, definitely
		Yes, if the price is low
		Not sure
		Probably not
		Definitely not

Notes. ⋆ = compared with log-derived analogues in the analyses (Q7–Q17). ◇ = multiple answers allowed (Q12–Q17).

C

Reproducibility and Code Structure

Repository

All analysis and generation scripts are available at: github.com/QV25/thesis-repo. The reproducibility repository contains only the source code required to regenerate analyses and figures. Top-level helpers and all files in scripts/ are tracked; data, caches, logs and exports are intentionally ignored. See the README for environment setup and script entry points. And the repository map shown below.

Table C.1: Prototypes

Family / Label	Prototype phrases
Q12 Main tasks	
Writing & professional communication	Writing an email, report or other professional text
Brainstorming & personal ideas – fun	Brainstorming ideas or asking fun creative questions
Coding – programming help	Getting help with coding or programming
Language practice or translation	Practising a foreign language or translating text
Study revision – exam prep	Studying, revising or preparing for an exam
Other	Any other kind of task
Q13 Writing subtasks	
Outline	Creating an outline for a presentation or slide deck
Draft	Drafting a complete email, letter or report for me
Proof-read / tone	Proof-reading my text and correcting tone or grammar
Summarise	Summarising articles, sources or meeting notes
Rewrite for audience	Rewriting the same text for different audiences
Other / no-choice	Any other use, or I did not choose Writing & professional communication
Q14 Brainstorming subtasks	
Academic ideas	Brainstorming academic or research ideas and paper topics
Business / product / marketing	Brainstorming business plans, product or marketing concepts
Creative role-play / jokes / stories	Creative role-play, jokes or storytelling with ChatGPT
Hypotheticals	Asking hypothetical what-if or alternate reality questions
Recommendations	Requesting recommendations for books, movies or music
Trivia	Asking trivia or general knowledge questions for fun
Other / no-choice	Any other use, or I did not choose Brainstorming & personal ideas
Q15 Coding subtasks	
Generate code	Generate fresh code snippets or function templates for me
Debug	Debug my existing code and fix errors
Explain	Explain what a piece of code does or clarify a concept
Convert	Convert code from one programming language to another
Unit tests	Write sample unit tests for my functions
Other / no-choice	Any other use, or I did not choose Coding / programming help
Q16 Language / translation subtasks	
Translate	Translate an entire paragraph or document from one language into another
Improve style	Improve my grammar or writing style in the target language
Vocabulary	Give me vocabulary drills or word lists to study
Role-play	Do a conversational role-play so I can practise dialogue
Pronunciation	Help with pronunciation or phonetic transcription
Other / no-choice	Any other use, or I did not choose Language practice or translation
Q17 Study / exam subtasks	
Summarise notes	Summarise my lecture notes or textbook chapter concisely
Practice questions	Generate practice questions or quizzes for my exam
Explain concept	Explain a difficult concept to me in simple terms
Mnemonics	Create mnemonics or other memory aids for key facts
Flashcards	Help me review flashcards or key terms for the test
Other / no-choice	Any other use, or I did not choose Study revision / exam prep

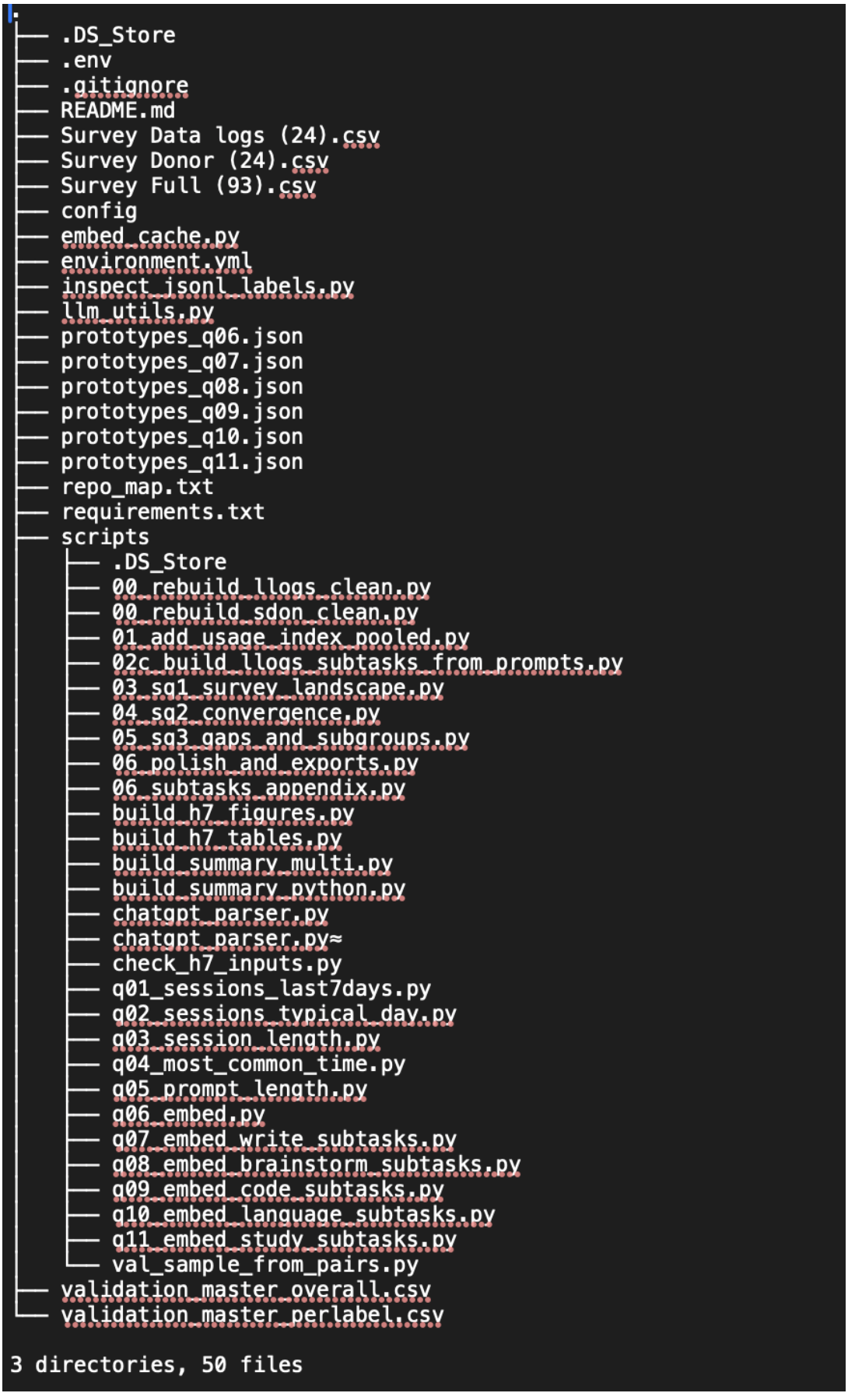

Figure C.1: GitHub repository map

D

Tables for Chapters 6 and 7

This appendix collects all tables referenced in Chapters 6 and 7. Each input file defines its own caption and label, so cross-references from the main text resolve to the entries listed here.

D.1 §6 Results

Table D.1: Q12 co-selection (Jaccard index)

(n) Category	1	2	3	4	5
1) q12_writing_and_professional_communication	1.000	0.506	0.273	0.343	0.059
2) q12_brainstorming_and_personal_ideas_fun	0.506	1.000	0.233	0.250	0.088
3) q12_coding_programming_help	0.273	0.233	1.000	0.333	0.111
4) q12_language_practice_or_translation	0.343	0.250	0.333	1.000	0.091
5) q12_other	0.059	0.088	0.111	0.091	1.000

Table D.3: SQ1 component loadings (oblimin rotation) and communalities.

	Dim1	Dim2	Dim3
usage_index_survey	0.18	**-0.60**	0.08
Q11_score	**-0.55**	-0.01	0.22
task_breadth_main	0.13	-0.14	**-0.64**
subtask_breadth_wri	**-0.39**	-0.27	**-0.40**
subtask_breadth_bra	-0.13	**-0.64**	-0.11
subtask_breadth_cod	**-0.46**	0.18	-0.19
subtask_breadth_lan	0.01	**0.33**	**-0.55**
Q18_mid	**-0.51**	-0.02	0.12

Table D.4: Association matrix (absolute ρ, V, and η).

(n) Variabele	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1) Q7_mid	1.000	0.924	0.888	0.783	0.929	0.913	0.963	0.867	0.889	0.888	0.918	0.952	0.537	0.323	0.267	0.444	0.227	0.240	0.222	0.158	0.188	0.192	0.410	0.051	0.136
2) Q8_mid	0.924	1.000	0.893	0.771	0.902	0.909	0.935	0.940	0.933	0.854	0.933	0.885	0.490	0.140	0.283	0.315	0.242	0.138	0.138	0.143	0.145	0.145	0.281	0.037	0.056
3) Q9_mid	0.888	0.893	1.000	0.858	0.883	0.924	0.959	0.955	0.873	0.912	0.944	0.886	0.360	0.331	0.449	0.296	0.322	0.027	0.063	0.281	0.104	0.000	0.253	0.031	0.017
4) Q11_score	0.783	0.771	0.858	1.000	0.804	0.882	0.797	0.816	0.958	0.855	0.808	0.858	0.224	0.133	0.316	0.240	0.258	0.015	0.364	1.000	0.094	0.039	0.091	0.156	0.089
5) task_breadth_main	0.929	0.902	0.883	0.804	1.000	0.921	0.885	0.879	0.864	0.921	0.891	0.932	0.434	0.413	0.376	0.463	0.378	0.354	0.479	0.499	0.538	0.490	0.643	0.600	0.203
6) subtask_breadth_wri	0.913	0.909	0.924	0.882	0.921	1.000	0.911	0.951	0.903	0.857	0.909	0.917	0.428	0.326	0.191	0.265	0.206	0.096	0.064	0.471	0.000	0.195	0.247	0.123	0.039
7) subtask_breadth_bra	0.963	0.935	0.959	0.797	0.885	0.911	1.000	0.942	0.843	0.857	0.938	0.931	0.265	0.263	0.176	0.315	0.234	0.160	0.236	0.259	0.050	0.000	0.149	0.103	0.066
8) subtask_breadth_cod	0.867	0.940	0.955	0.816	0.879	0.951	0.942	1.000	0.925	0.917	0.939	0.862	0.405	0.199	0.087	0.320	0.428	0.295	0.052	0.398	0.248	0.101	0.000	0.021	0.146
9) subtask_breadth_lan	0.889	0.933	0.873	0.958	0.864	0.903	0.843	0.925	1.000	0.774	0.884	0.816	0.272	0.599	0.157	0.360	0.482	0.312	0.346	0.286	0.194	0.522	0.425	0.000	0.013
10) subtask_breadth_stu	0.888	0.854	0.912	0.855	0.921	0.857	0.857	0.917	0.774	1.000	0.823	0.895	0.049	0.405	0.435	0.444	0.127	0.253	0.056	0.537	0.474	0.026	0.105	0.222	0.053
11) usage_index_survey	0.918	0.933	0.944	0.808	0.891	0.909	0.938	0.939	0.884	0.823	1.000	0.883	0.134	0.067	0.128	0.432	0.112	0.098	0.073	0.070	0.082	0.035	0.011	0.132	0.017
12) Q18_mid	0.952	0.885	0.886	0.858	0.932	0.917	0.931	0.862	0.816	0.895	0.883	1.000	0.250	0.333	0.352	0.243	0.238	0.013	0.385	0.316	0.225	0.274	0.121	0.044	0.154
13) Q3	0.537	0.490	0.360	0.224	0.434	0.428	0.265	0.405	0.272	0.049	0.134	0.250	1.000	0.597	0.590	0.634	0.599	0.713	0.436	0.464	0.341	0.264	0.297	0.186	0.229
14) Q4	0.323	0.140	0.331	0.133	0.413	0.326	0.263	0.199	0.599	0.405	0.067	0.333	0.597	1.000	0.610	0.627	0.588	0.713	0.473	0.474	0.359	0.250	0.251	0.151	0.135
15) Q5	0.267	0.283	0.449	0.316	0.376	0.191	0.176	0.087	0.157	0.435	0.128	0.352	0.590	0.610	1.000	0.547	0.636	0.726	0.458	0.411	0.413	0.286	0.230	0.236	0.091
16) Q6	0.444	0.315	0.296	0.240	0.463	0.265	0.315	0.320	0.360	0.444	0.432	0.243	0.634	0.627	0.547	1.000	0.490	0.759	0.489	0.404	0.381	0.364	0.493	0.228	0.254
17) Q2	0.227	0.242	0.322	0.258	0.378	0.206	0.234	0.428	0.482	0.127	0.112	0.238	0.599	0.588	0.636	0.490	1.000	0.716	0.490	0.407	0.373	0.303	0.262	0.273	0.234
18) Q1	0.240	0.138	0.027	0.015	0.354	0.096	0.160	0.295	0.312	0.253	0.098	0.013	0.713	0.713	0.726	0.759	0.716	1.000	0.533	0.558	0.313	0.268	0.187	0.157	0.096
19) Q10	0.222	0.138	0.063	0.364	0.479	0.064	0.236	0.052	0.346	0.056	0.073	0.385	0.436	0.473	0.458	0.489	0.490	0.533	1.000	0.615	0.426	0.343	0.189	0.184	0.116
20) Q11_band	0.158	0.143	0.281	1.000	0.499	0.471	0.259	0.398	0.286	0.537	0.070	0.316	0.464	0.474	0.411	0.404	0.407	0.558	0.615	1.000	0.446	0.341	0.379	0.343	0.128
21) q12_writing_and professional_communication	0.188	0.145	0.104	0.094	0.538	0.000	0.050	0.248	0.194	0.474	0.082	0.225	0.341	0.359	0.413	0.381	0.373	0.313	0.426	0.446	1.000	0.144	0.208	0.239	0.163
22) q12_brainstorming_and personal_ideas_fun	0.192	0.145	0.000	0.039	0.490	0.195	0.000	0.101	0.522	0.026	0.035	0.274	0.264	0.250	0.286	0.364	0.303	0.268	0.343	0.341	0.144	1.000	0.106	0.029	0.009
23) q12_coding_programming_help	0.410	0.281	0.253	0.091	0.643	0.247	0.149	0.000	0.425	0.105	0.011	0.121	0.297	0.251	0.230	0.493	0.262	0.187	0.189	0.379	0.208	0.106	1.000	0.334	0.084
24) q12_language_practice_or translation	0.051	0.037	0.031	0.156	0.600	0.123	0.103	0.021	0.000	0.222	0.132	0.044	0.186	0.151	0.236	0.228	0.273	0.157	0.184	0.343	0.239	0.029	0.334	1.000	0.031
25) q12_other	0.136	0.056	0.017	0.089	0.203	0.039	0.066	0.146	0.013	0.053	0.017	0.154	0.229	0.135	0.091	0.254	0.234	0.096	0.116	0.128	0.163	0.009	0.084	0.031	1.000

Table D.6: SQ1 OLS models — Dimension 1.

Term	Estimate [95% CI]	Partial R^2
const	-0.20 [-0.20, -0.20]	1.000
Q3 (Plan): Free	0.42 [0.03, 0.81]	0.056
Q3 (Plan): Plus (20 euro)	-0.19 [-0.61, 0.23]	0.011
Q3 (Plan): Pro (200 euro)	0.17 [-0.37, 0.72]	0.005
Q4 (Device): Laptop / desktop	-0.09 [-0.50, 0.31]	0.003
Q4 (Device): Mixed equally	0.12 [-0.42, 0.66]	0.003
Q4 (Device): Smartphone	0.37 [-0.12, 0.85]	0.029
Q5 (Status): Both student and working	-0.15 [-0.99, 0.70]	0.002
Q5 (Status): Other	0.54 [-1.13, 2.22]	0.005
Q5 (Status): Student	0.01 [-0.78, 0.80]	0.000
Q5 (Status): Working	-0.01 [-0.61, 0.60]	0.000
Q6 (Field): Business / economics	0.08 [-0.70, 0.86]	0.001
Q6 (Field): Creative arts / media	0.70 [-0.24, 1.64]	0.028
Q6 (Field): Humanities / social sciences	-0.25 [-1.12, 0.63]	0.004
Q6 (Field): Other	-0.66 [-1.55, 0.22]	0.029
Q6 (Field): Prefer not to say	0.56 [-3.07, 4.19]	0.001
Q6 (Field): STEM (science, tech, engineering, maths)	-0.03 [-0.85, 0.79]	0.000
Q2 (Age): 18-24	-0.28 [-1.27, 0.71]	0.004
Q2 (Age): 25-34	-0.24 [-1.22, 0.75]	0.003
Q2 (Age): 35-44	-0.39 [-1.42, 0.63]	0.008
Q2 (Age): 45+	0.30 [-0.88, 1.47]	0.003
Q2 (Age): <18	1.01 [-3.03, 5.05]	0.003
Q1 (Gender): Man	0.20 [-0.24, 0.64]	0.010
Q1 (Gender): Woman	0.20 [-0.27, 0.67]	0.009

Table D.7: SQ1 OLS models — Dimension 2.

Term	Estimate [95% CI]	Partial R^2
const	0.24 [0.24, 0.24]	1.000
Q3 (Plan): Free	0.19 [-0.17, 0.55]	0.014
Q3 (Plan): Plus (20 euro)	-0.32 [-0.75, 0.10]	0.030
Q3 (Plan): Pro (200 euro)	-0.24 [-0.83, 0.34]	0.009
Q4 (Device): Laptop / desktop	-0.47 [-0.82, -0.13]	0.089
Q4 (Device): Mixed equally	0.33 [-0.33, 0.99]	0.013
Q4 (Device): Smartphone	-0.24 [-0.70, 0.23]	0.013
Q5 (Status): Both student and working	-0.39 [-1.23, 0.45]	0.011
Q5 (Status): Other	0.05 [-1.51, 1.61]	0.000
Q5 (Status): Student	-0.23 [-0.84, 0.37]	0.008
Q5 (Status): Working	0.19 [-0.42, 0.80]	0.005
Q6 (Field): Business / economics	0.36 [-0.15, 0.86]	0.025
Q6 (Field): Creative arts / media	0.13 [-0.53, 0.80]	0.002
Q6 (Field): Humanities / social sciences	0.93 [0.25, 1.60]	0.089
Q6 (Field): Other	0.13 [-0.35, 0.62]	0.004
Q6 (Field): Prefer not to say	-2.15 [-3.30, -1.01]	0.156
Q6 (Field): STEM (science, tech, engineering, maths)	0.22 [-0.33, 0.78]	0.008
Q2 (Age): 18-24	0.18 [-0.29, 0.65]	0.008
Q2 (Age): 25-34	-0.00 [-0.57, 0.56]	0.000
Q2 (Age): 35-44	-0.66 [-1.46, 0.14]	0.034
Q2 (Age): 45+	-0.21 [-0.99, 0.57]	0.004
Q2 (Age): <18	0.31 [-1.24, 1.87]	0.002
Q1 (Gender): Man	-0.22 [-0.54, 0.11]	0.023
Q1 (Gender): Woman	-0.16 [-0.58, 0.26]	0.007

Table D.8: SQ1 OLS models — Dimension 3.

Term	Estimate [95% CI]	Partial R^2
const	1.10 [1.10, 1.10]	1.000
Q3 (Plan): Free	0.11 [-0.20, 0.43]	0.006
Q3 (Plan): Plus (20 euro)	-0.54 [-0.94, -0.13]	0.084
Q3 (Plan): Pro (200 euro)	-0.29 [-0.69, 0.12]	0.025
Q4 (Device): Laptop / desktop	-0.36 [-0.72, 0.01]	0.048
Q4 (Device): Mixed equally	-0.22 [-1.02, 0.58]	0.004
Q4 (Device): Smartphone	-0.14 [-0.58, 0.31]	0.005
Q5 (Status): Both student and working	0.06 [-0.55, 0.67]	0.001
Q5 (Status): Other	-0.36 [-1.22, 0.51]	0.009
Q5 (Status): Student	-0.19 [-0.71, 0.32]	0.007
Q5 (Status): Working	-0.22 [-0.66, 0.21]	0.013
Q6 (Field): Business / economics	-0.21 [-0.70, 0.28]	0.010
Q6 (Field): Creative arts / media	-0.29 [-0.96, 0.38]	0.010
Q6 (Field): Humanities / social sciences	0.18 [-0.31, 0.67]	0.007
Q6 (Field): Other	-0.07 [-0.43, 0.30]	0.002
Q6 (Field): Prefer not to say	0.25 [-0.31, 0.81]	0.010
Q6 (Field): STEM	-0.57 [-1.11, -0.03]	0.055
Q2 (Age): 18-24	0.10 [-0.27, 0.47]	0.004
Q2 (Age): 25-34	-0.25 [-0.65, 0.15]	0.019
Q2 (Age): 35-44	0.05 [-0.58, 0.68]	0.000
Q2 (Age): 45+	0.14 [-0.40, 0.69]	0.004
Q2 (Age): <18	-0.75 [-1.64, 0.14]	0.036
Q1 (Gender): Man	-0.11 [-0.41, 0.19]	0.007
Q1 (Gender): Woman	-0.60 [-0.96, -0.23]	0.123

Table D.9: Numeric contrasts: HL shifts, KS, Wasserstein, n and FDR.

component	label	effect	l95	u95	p	cliffs_delta	ks	w1	n_Sdon	n_Llogs	q_FDR
Q7_mid	Sessions per week	2.500	0.000	4.000	0.940	0.262	0.208	1.688	24.000	24.000	0.940
Q8_mid	Sessions per day	0.000	0.000	1.500	0.800	0.260	0.292	0.708	24.000	24.000	0.940
Q9_mid	Minutes per session	0.000	−12.500	12.500	0.185	−0.042	0.458	10.506	24.000	14.000	0.370
usage_index	Usage index (pooled z)	0.508	0.000	0.894	0.078	0.316	0.333	0.489	24.000	24.000	0.310

Table D.10: Q10 timing: shares (Wilson), Δp (Newcombe), Cramér's V, n, FDR.

category	p_Sdon	l_Sdon	u_Sdon	p_Llogs	l_Llogs	u_Llogs	deltap	l95	u95	p	n_Sdon	n_Llogs	q_FDR	cramers_V
During work / study hours	0.625	0.427	0.788	0.708	0.508	0.851	−0.083	−0.424	0.280	0.540	24.000	24.000	0.810	0.093
Evenings	0.042	0.007	0.202	0.042	0.007	0.202	0.000	−0.195	0.195	1.000	24.000	24.000	1.000	0.093
Anytime throughout the day	0.333	0.180	0.533	0.250	0.120	0.449	0.083	−0.269	0.413	0.525	24.000	24.000	0.810	0.093

Table D.11: Q11 prompt length: shares and Δp with FDR; table-level association.

category	p_Sdon	l_Sdon	u_Sdon	p_Llogs	l_Llogs	u_Llogs	deltap	l95	u95	p	n_Sdon	n_Llogs	q_FDR	cramers_V
One short sentence	0.250	0.120	0.449	0.667	0.467	0.820	−0.417	−0.700	−0.018	0.004	24.000	24.000	0.008	0.655
A short paragraph	0.583	0.388	0.755	0.042	0.007	0.202	0.542	0.186	0.748	0.000	24.000	24.000	0.000	0.655
Multiple paragraphs	0.083	0.023	0.258	0.000	0.000	0.138	0.083	−0.115	0.258	0.149	24.000	24.000	0.149	0.655
Varies too much to say	0.083	0.023	0.258	0.292	0.149	0.492	−0.208	−0.469	0.109	0.064	24.000	24.000	0.086	0.655

Table D.12: Q11 linear-by-linear trend (logit slope with 95% CI).

coef	l95	u95	n
3.628822	1.405095	5.852548	39

Table D.14: Q12 family prevalence by cohort with Δp and FDR.

family	Sdon_p	Sdon_l	Sdon_u	Llogs_p	Llogs_l	Llogs_u	deltap	l95	u95	p	q_FDR
Q12: Writing & communication	0.875	0.690	0.957	0.750	0.551	0.880	0.125	−0.190	0.406	0.267	0.334
Q12: Brainstorming / fun	0.667	0.467	0.820	0.667	0.467	0.820	0.000	−0.353	0.353	1.000	1.000
Q12: Coding / programming	0.250	0.120	0.449	0.625	0.427	0.788	−0.375	−0.668	0.022	0.009	0.015
Q12: Language / translation	0.292	0.149	0.492	0.708	0.508	0.851	−0.417	−0.702	−0.017	0.004	0.010
Q12: Other	0.083	0.023	0.258	1.000	0.862	1.000	−0.917	−0.977	−0.604	0.000	0.000

Table D.15: Task-breadth by cohort: medians, HL shift, MWU, Cliff's δ.

component	median_Sdon	median_Llogs	HL	l95	u95	p	cliffs_delta
Task breadth (Q12 families)	2.0	4.0	-2.0	-2.0	-2.0	0.000002	-0.71875

Table D.17: Task-pattern alignment: Spearman ρ (95% CI) and Hellinger distance.

Table D.18: T6 7 pattern ci

spearman_rho	l95	u95	hellinger
-0.1	-1.0	1.0	0.585734

Table D.19: Survey-only model for usage index (OLS, HC3).

Term	Std. beta	HC3 s.e.
const	-0.053	0.089
Q3_Plus	0.517	0.116
Q3_Pro	0.276	0.086
Q4_Mixed equally	-0.024	0.123
Q4_Smartphone	-0.057	0.095
Q5_Both student and working	0.240	0.132
Q5_Other	-0.070	0.099
Q5_Student	0.085	0.147
Q6_Business / economics	-0.094	0.140
Q6_Creative arts / media	0.067	0.124
Q6_Humanities / social sciences	-0.139	0.112
Q6_Other	-0.180	0.121
Q6_Prefer not to say	-0.007	0.042
Q2_18–24	-0.102	0.101
Q2_18–24 (alt)	0.011	0.019
Q2_25–34	0.126	0.069
Q2_35–44	0.023	0.058
Q2_45+	-0.048	0.069
Q2_<18	-0.065	0.043

Table D.20: Ordered/logit model for Q11 “A short paragraph”.

Term	Std. beta (LPM)	HC3 s.e.
const	-0.000	0.111
Q3_Plus	0.052	0.139
Q3_Pro	-0.084	0.128
Q4_Mixed equally	0.072	0.142
Q4_Smartphone	0.252	0.146
Q5_Both student and working	0.037	0.149
Q5_Other	-0.340	0.155
Q5_Student	-0.173	0.194
Q6_Business / economics	0.003	0.156
Q6_Creative arts / media	-0.106	0.150
Q6_Humanities / social sciences	-0.019	0.128
Q6_Other	0.081	0.158
Q6_Prefer not to say	-0.185	0.049
Q2_18–24	0.154	0.137
Q2_18–24	0.062	0.140
Q2_25–34	-0.007	0.094
Q2_35–44	-0.133	0.098
Q2_45+	-0.052	0.132
Q2_<18	-0.184	0.060

Table D.21: Logit model for Q12 Coding/programming (selected vs not).

Term	Std. beta (LPM)	HC3 s.e.
const	0.000	0.099
Q3_Plus	0.199	0.127
Q3_Pro	0.052	0.112
Q4_Mixed equally	-0.111	0.136
Q4_Smartphone	-0.054	0.106
Q5_Both student and working	0.144	0.126
Q5_Other	0.081	0.084
Q5_Student	0.079	0.161
Q6_Business / economics	-0.363	0.160
Q6_Creative arts / media	-0.144	0.153
Q6_Humanities / social sciences	-0.412	0.101
Q6_Other	-0.424	0.119
Q6_Prefer not to say	-0.080	0.043
Q2_18–24	-0.042	0.092
Q2_18–24	-0.045	0.083
Q2_25–34	0.126	0.075
Q2_35–44	-0.081	0.072
Q2_45+	-0.009	0.067
Q2_<18	-0.020	0.043

Table D.22: Logit model for Q12 Language/translation (selected vs not).

Term	Std. beta (LPM)	HC3 s.e.
const	0.000	0.117
Q3_Plus	0.080	0.142
Q3_Pro	-0.045	0.141
Q4_Mixed equally	-0.043	0.133
Q4_Smartphone	-0.073	0.144
Q5_Both student and working	-0.194	0.164
Q5_Other	0.118	0.317
Q5_Student	-0.208	0.210
Q6_Business / economics	-0.034	0.166
Q6_Creative arts / media	-0.111	0.138
Q6_Humanities / social sciences	0.073	0.153
Q6_Other	-0.182	0.167
Q6_Prefer not to say	-0.075	0.054
Q2_18–24	-0.003	0.153
Q2_18–24	-0.186	0.261
Q2_25–34	0.102	0.103
Q2_35–44	-0.042	0.111
Q2_45+	0.003	0.121
Q2_<18	0.032	0.059

Table D.23: Representativeness: Sdon vs Survey full (Cramér's V by item).

Table D.24: T6 A repr V

question	cramers_V	df	n
Q1	0.111463	2	117
Q2	0.186162	5	117
Q3	0.169970	3	117
Q4	0.164008	3	117
Q5	0.180173	4	117
Q6	0.154392	6	117
Q18	0.210696	5	117
Q19	0.281875	5	117
Q20	0.292352	5	117

Table D.25: Sensitivity: alternative midpoints and episode-split checks.

Measure	HLorig [95% CI]	HLsens [95% CI]	Stable?
Q7 (12→15)	2.50 [0.00, 4.00]	2.50 [0.00, 6.50]	yes
Q8 (6→7)	0.00 [0.00, 1.50]	0.00 [0.00, 1.50]	yes
Q9 (>60: 75→90)	0.00 [0.00, 22.50]	0.00 [0.00, 22.50]	yes

D.2 §7 Validation

Label	TP	FP	FN	Precision	Recall	F1	Support
WRI	8	2	5	0.80	0.62	0.70	13
BRA	3	10	0	0.23	1.00	0.38	3
COD	8	12	0	0.40	1.00	0.57	8
LAN	4	4	1	0.50	0.80	0.62	5
STU	15	1	12	0.94	0.56	0.70	27
OTH	25	8	20	0.76	0.56	0.64	45

Table D.26: Q12 per-label metrics. Codes per Table 4.4

Label	TP	FP	FN	Precision	Recall	F1	Support
WRI	2	8	0	0.20	1.00	0.33	2
BRA	3	11	6	0.21	0.33	0.26	9
COD	2	4	2	0.33	0.50	0.40	4
LAN	3	16	1	0.16	0.75	0.26	4
STU	5	11	0	0.31	1.00	0.48	5
OTH	34	1	42	0.97	0.45	0.61	76

Table D.27: Q13 per-label metrics. Codes per Table 4.4

Label	TP	FP	FN	Precision	Recall	F1	Support
WRI	4	4	2	0.50	0.67	0.57	6
BRA	1	12	2	0.08	0.33	0.13	3
COD	2	13	0	0.13	1.00	0.24	2
LAN	2	19	0	0.10	1.00	0.17	2
STU	1	2	0	0.33	1.00	0.50	1
OTH	32	4	47	0.89	0.41	0.56	79
TRI	3	1	5	0.75	0.38	0.50	8

Table D.28: Q14 per-label metrics. Codes per Table 4.4.

Label	TP	FP	FN	Precision	Recall	F1	Support
WRI	1	11	1	0.08	0.50	0.14	2
BRA	2	7	2	0.22	0.50	0.31	4
COD	3	21	1	0.13	0.75	0.21	4
LAN	1	3	0	0.25	1.00	0.40	1
STU	2	1	0	0.67	1.00	0.80	2
OTH	44	3	43	0.94	0.51	0.66	87

Table D.29: Q15 per-label metrics. Codes per Table 4.4.

Label	TP	FP	FN	Precision	Recall	F1	Support
WRI	1	9	2	0.10	0.33	0.15	3
BRA	1	6	2	0.14	0.33	0.20	3
COD	2	11	0	0.15	1.00	0.27	2
LAN	2	21	0	0.09	1.00	0.16	2
STU	2	6	0	0.25	1.00	0.40	2
OTH	38	1	50	0.97	0.43	0.60	88

Table D.30: Q16 per-label metrics. Codes per Table 4.4.

Label	TP	FP	FN	Precision	Recall	F1	Support
WRI	3	13	0	0.19	1.00	0.32	3
BRA	2	11	0	0.15	1.00	0.27	2
COD	20	7	2	0.74	0.91	0.82	22
LAN	2	15	0	0.12	1.00	0.21	2
OTH	27	0	44	1.00	0.38	0.55	71

Table D.31: Q17 per-label metrics. Codes per Table 4.4.

E

Figures for Chapters 6 and 7

This appendix reproduces the figures referenced in Chapters 6 and 7. Each included file or graphic provides its own caption and label; cross-references in the main text resolve here.

E.1 §6 Results

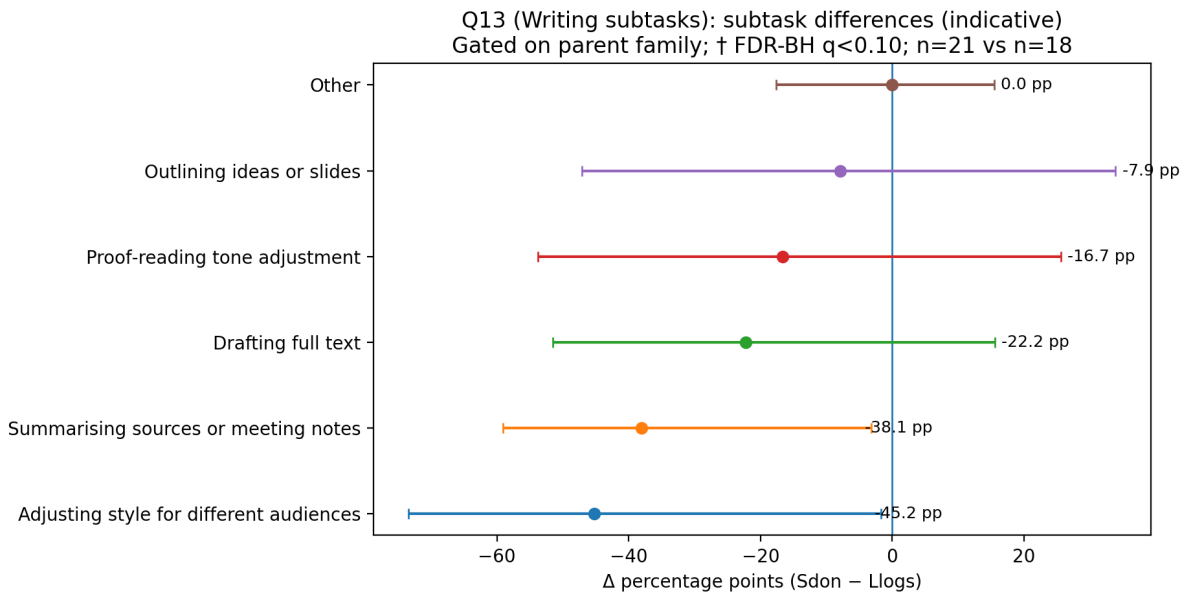

Figure E.1: Q13 (Writing) subtasks: Δp with 95% CIs (gated).

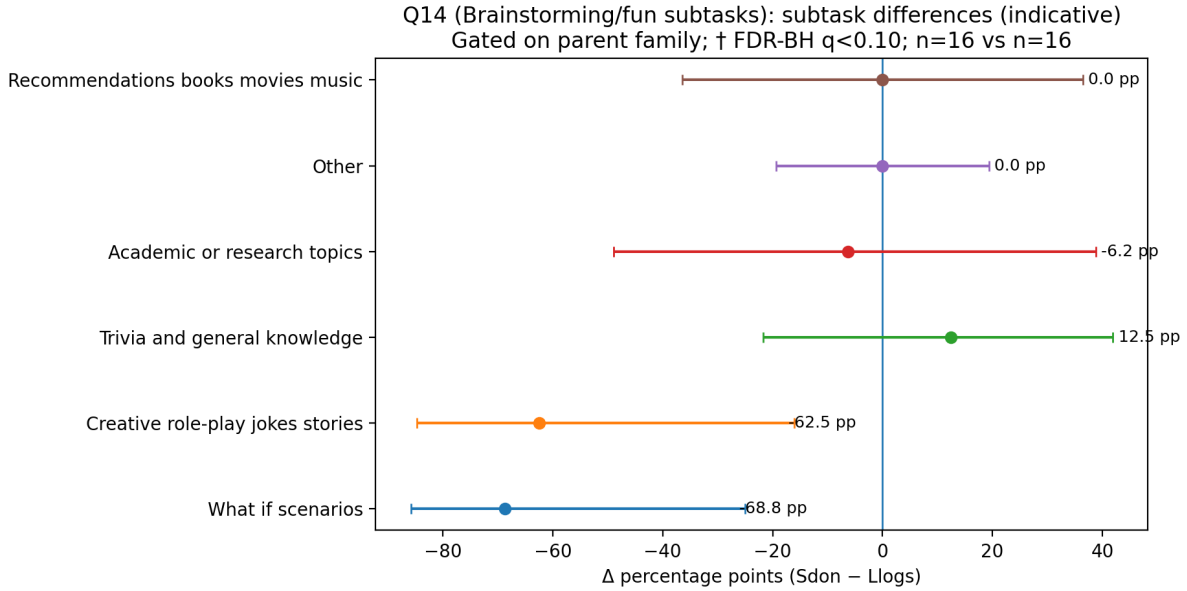

Figure E.2: Q14 (Brainstorming/fun) subtasks: Δp with 95% CIs (gated).

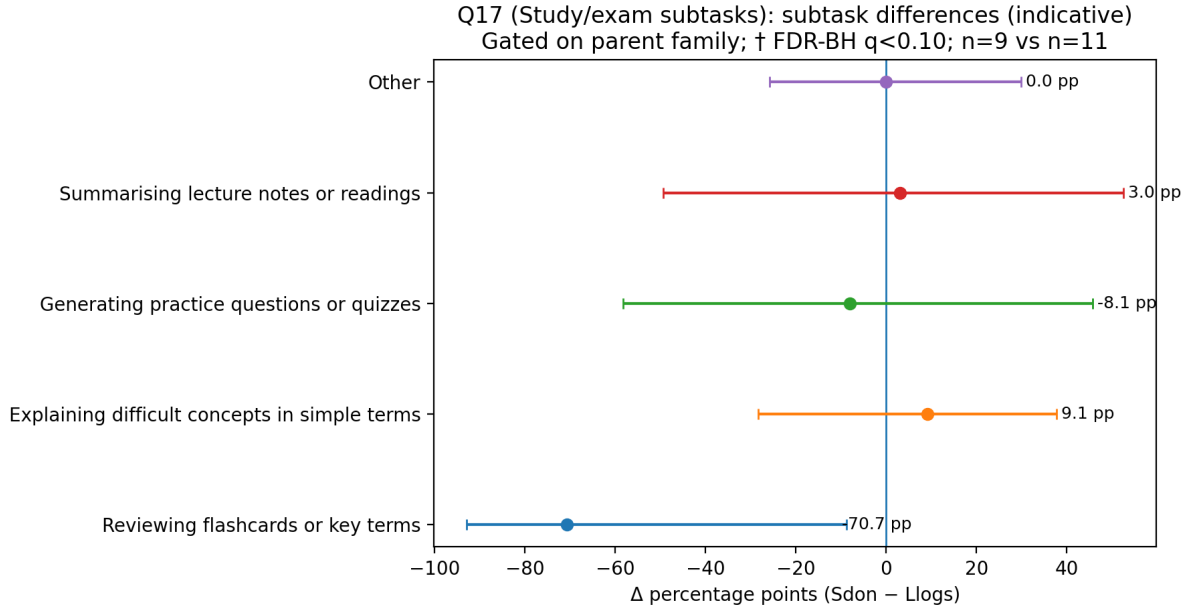

Figure E.3: Q17 (Study/exam) subtasks: Δp with 95% CIs (gated).

E.2 §7 Validation

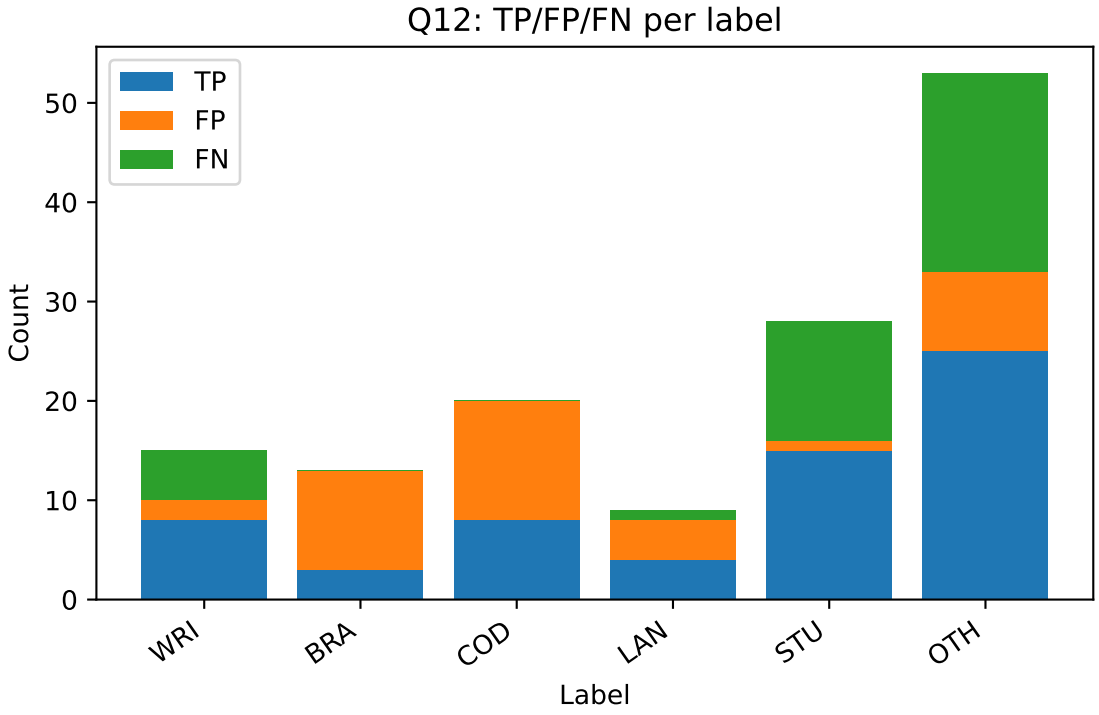

Figure E.4: Q12 TP/FP/FN per label. Codes per Table 4.4.

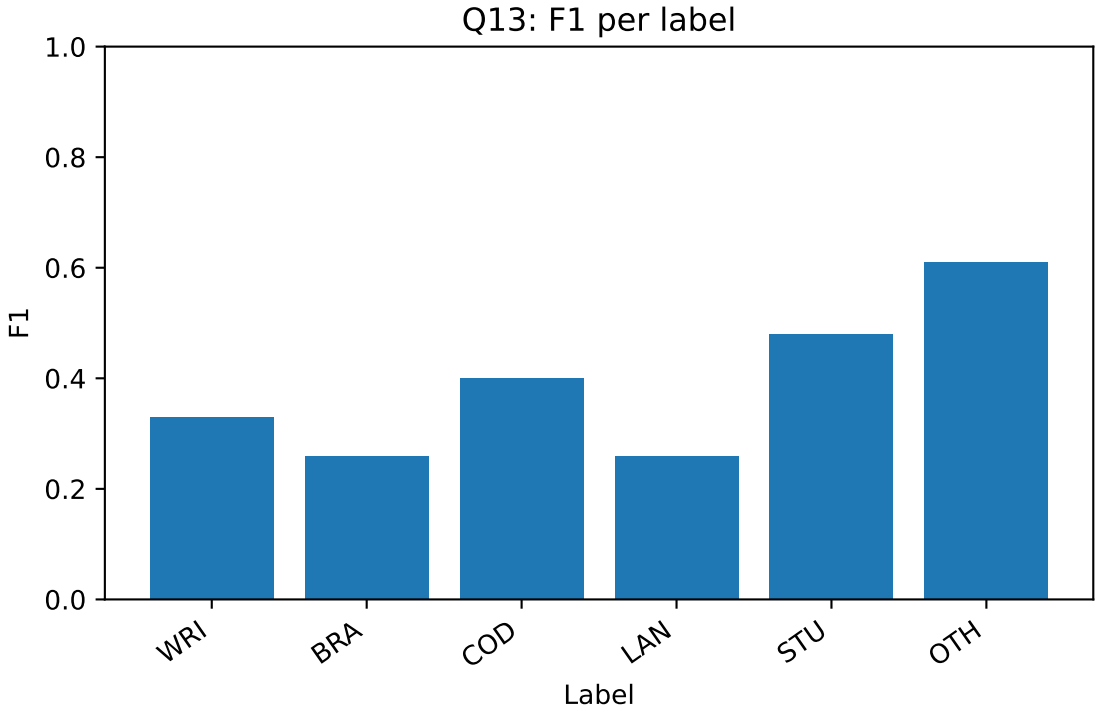

Figure E.5: Q13 per-label F1. Codes and label texts per Table 4.4.

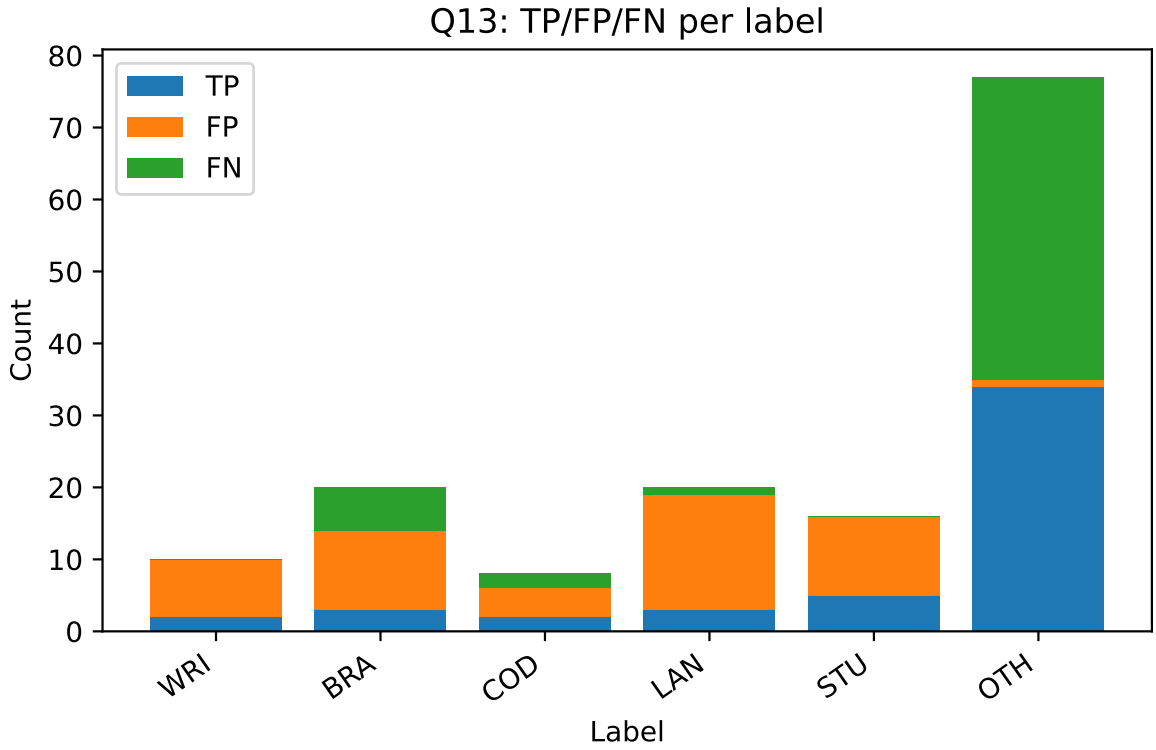

Figure E.6: Q13 TP/FP/FN per label. Codes per Table 4.4.

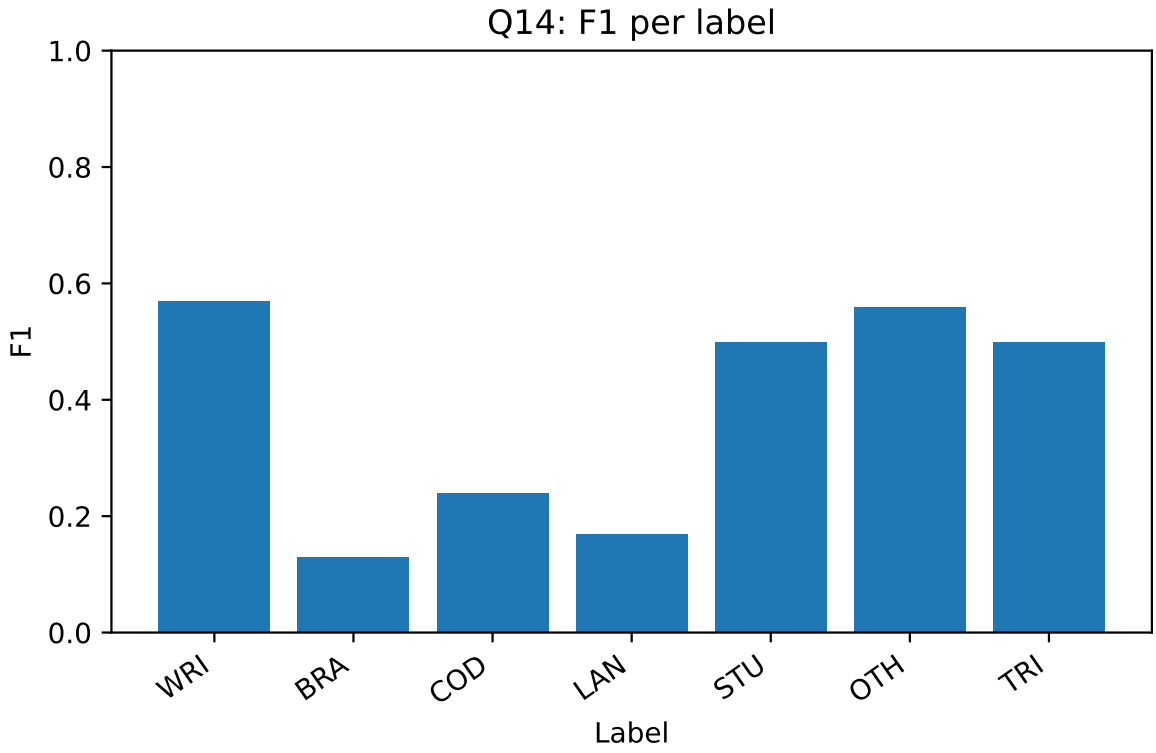

Figure E.7: Q14 per-label F1. Codes and label texts per Table 4.4.

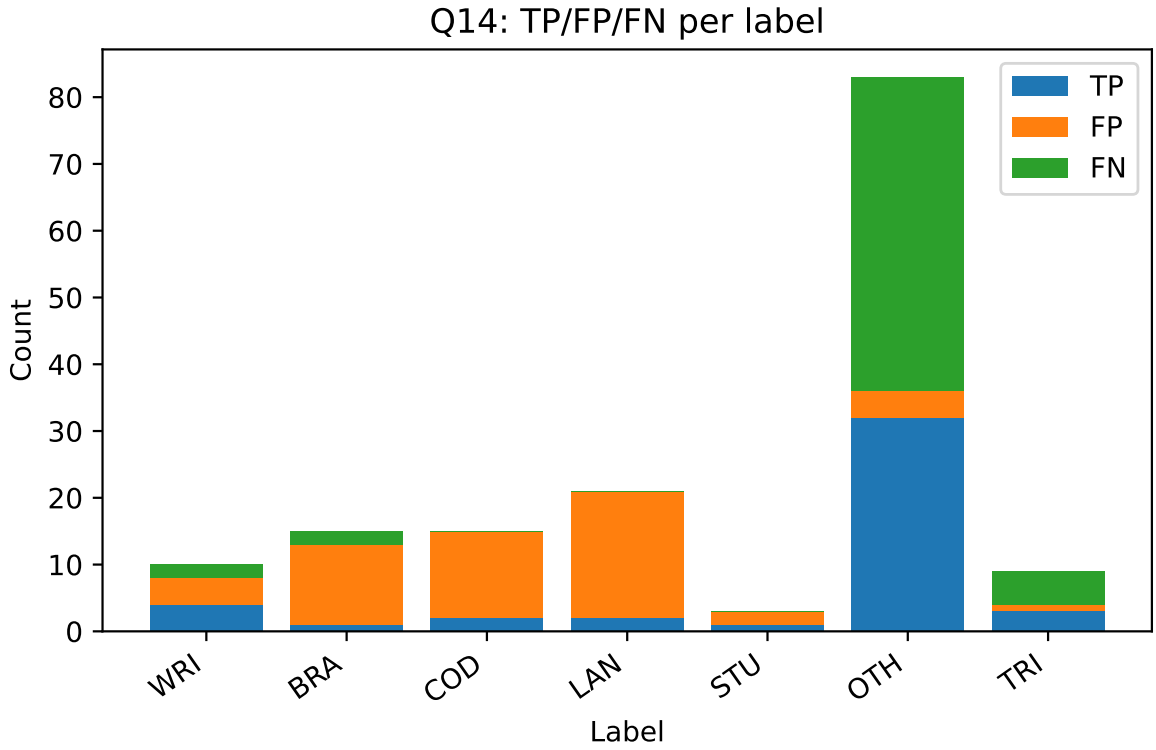

Figure E.8: Q14 TP/FP/FN per label. Codes per Table 4.4.

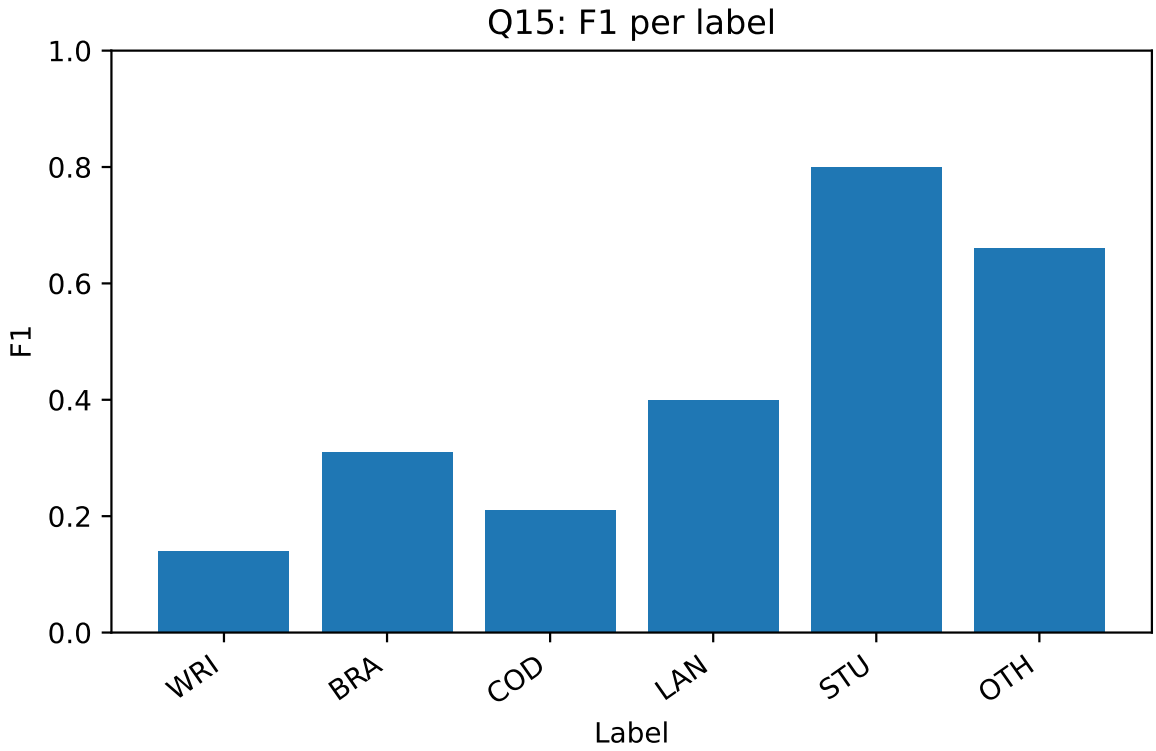

Figure E.9: Q15 per-label F1. Codes and label texts per Table 4.4.

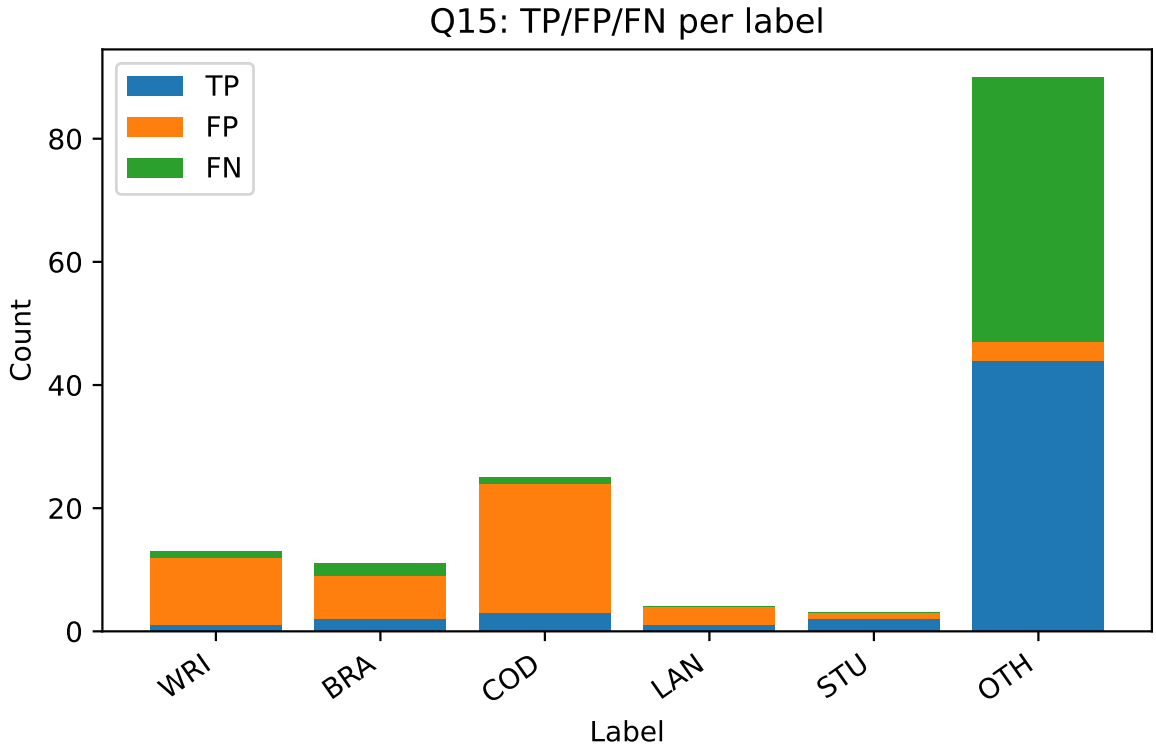

Figure E.10: Q15 TP/FP/FN per label. Codes per Table 4.4.

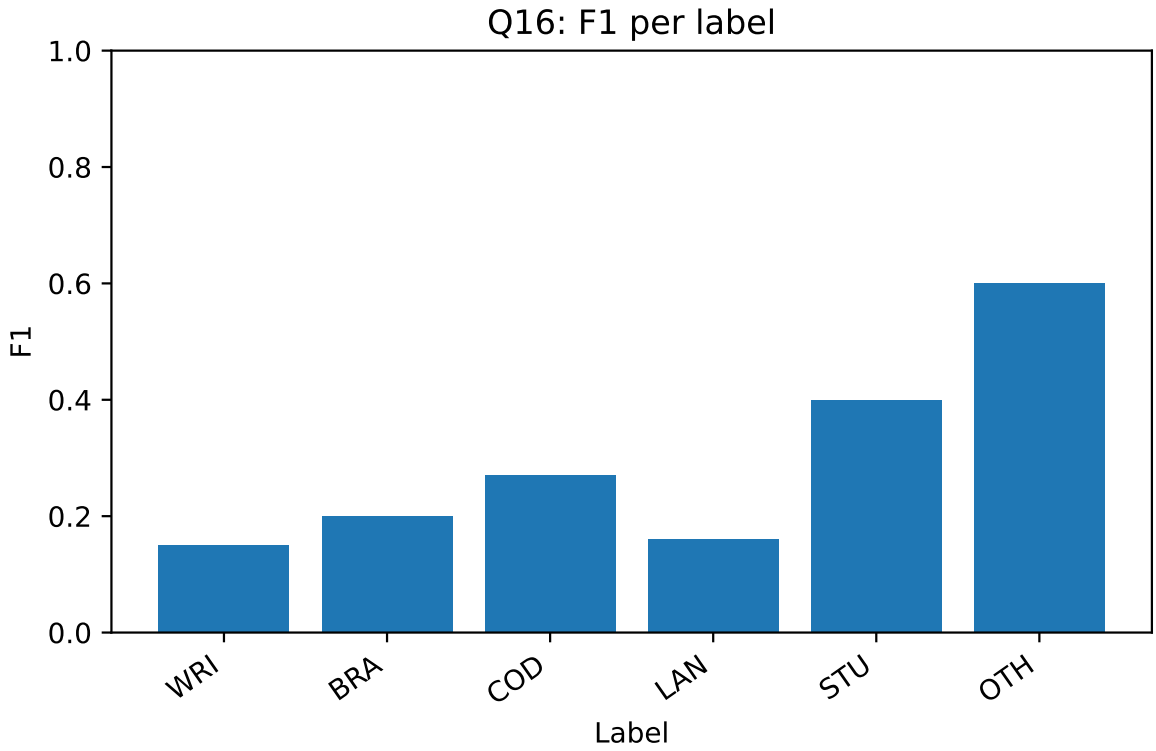

Figure E.11: Q16 per-label F1. Codes and label texts per Table 4.4.

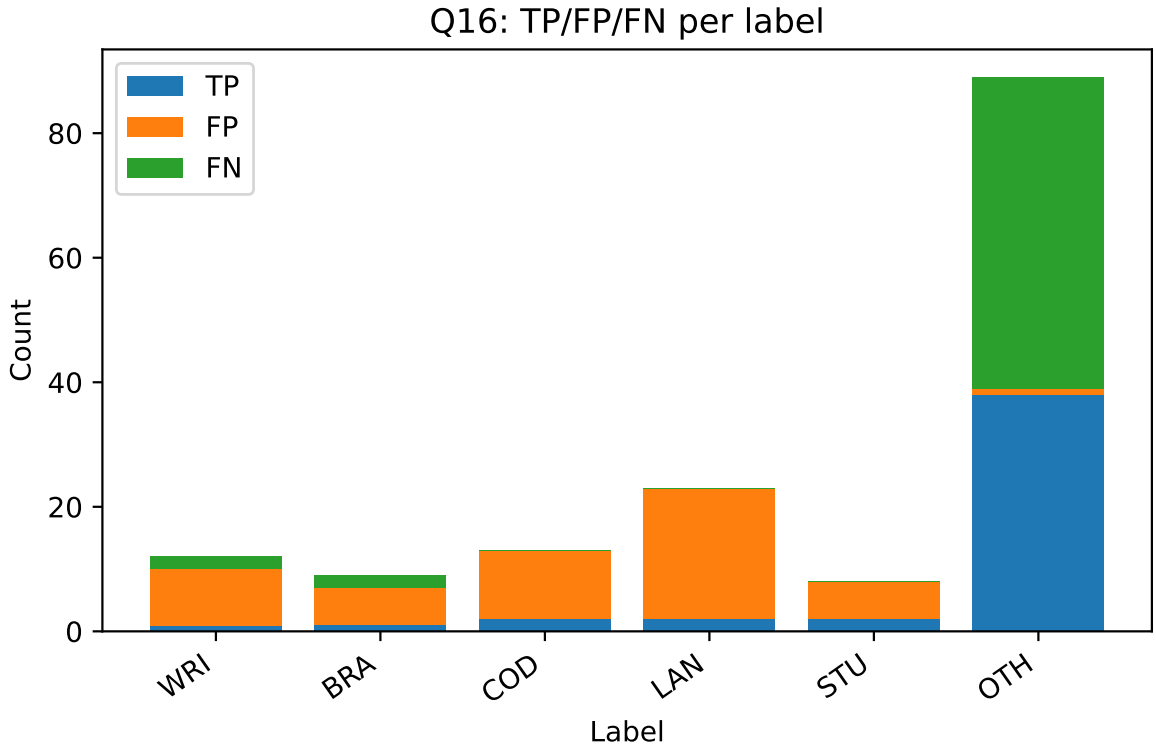

Figure E.12: Q16 TP/FP/FN per label. Codes per Table 4.4.

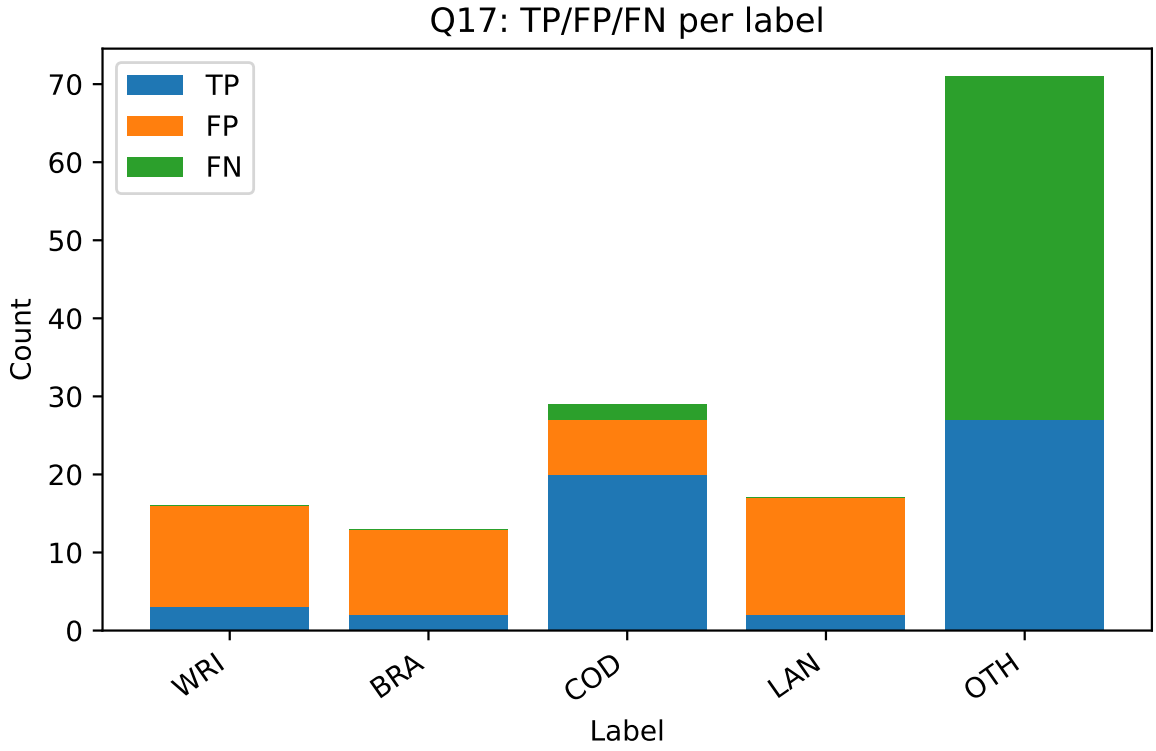

Figure E.13: Q17 TP/FP/FN per label. Codes per Table 4.4.

F

AI Use

This appendix documents, for transparency, how AI was used during the project. The guiding principle was simple: AI acted as a helper, not a ghostwriter. No donated logs or other sensitive materials were shared with third-party services. These choices align with the study's privacy-by-design protocol and donation workflow described in Chapters 3–4.

1. **Coding partner:** Used as a reviewer to check scripts, surface bugs and edge cases and suggest clearer or more efficient implementations. Suggestions were tested and kept only when they improved correctness or clarity; final code was written and validated manually.
2. **Source cross-checks:** After sources were first located and read, the assistant produced short bullet-point summaries and answered whether each source fit the intended use (measurement design, small-n estimation, privacy practice). Inclusion, phrasing and all citations relied on the source texts; any AI summaries were treated as a second opinion.
3. **Structure & wording:** Near submission, the assistant flagged unclear or repetitive sentences and proposed line-level edits. No sections were replaced wholesale; suggestions were reviewed and adapted to keep tone and meaning consistent with the rest of the thesis.
4. **Final consistency sweep:** The assistant checked for consistency across chapters (terminology, numbers, units, thresholds, figure/table references) and for general readability. Where it flagged mismatches, the text was harmonised manually; substantive claims and statistics were not altered.

In short, AI functioned as a careful reviewer of code and text. Ownership is taken for all analysis decisions, interpretations, final wording and any errors.