Session 3 Handout: Comparing and Extracting

FAST-02 Generative AI for Everyday Work · 7formations · Instructor: Wadood Chaudhary · All examples fictional

Why this session matters

Pulling facts out of messy documents (quotes, invoices, spec sheets, schedules) into a clean table is one of the most valuable things AI does at work. It is also where a single invented number can change a purchasing decision. This session teaches extraction that is fast *and* auditable, using three supplier quotes received by Bayside Precision Parts (fictional, Worcester County).

Key concepts

1. Define the table before you start

Decide the columns, units, and formats before the AI sees any document. A defined table:

- stops the AI from inventing its own structure for each document;
- forces consistent units (USD per lb, business days) and dates;
- makes a cell-by-cell audit possible.

Always include a **Source** column (document and line or page) so anyone can trace a value.

2. Blanks, not guesses

The most important instruction in an extraction prompt: **"If a value is not stated in the document, leave the cell blank."** Without it, models fill missing cells with plausible values: a "typical" lead time, an assumed certification, a standard warranty. These invented values are hard to spot because they look like everything else in the table.

3. One document at a time, labeled

Paste or attach one document at a time, labeled (Quote A, B, C). Mixed inputs invite the model to blend facts between suppliers. Remove anything the comparison doesn't need, such as bank details, signatures, and staff names, before extraction.

4. Images and PDFs

Most tools now read screenshots, scans, and PDFs. Useful, but the error rate rises: PDF tables can shift columns, scans misread digits (3.05 becomes 3.65), and handwriting is unreliable. OpenAI's own help center notes that ChatGPT "may not reliably extract exact values from image-based tables, scanned files, or files with complex visual layouts," and recommends spreadsheet or text files when exact values matter. Check line items and totals against the original image every time. The data rules apply to images too: a photo of a form with a name and account number is still personal information.

5. Audit every cell

Mark	Meaning	Example
Correct	Matches the source exactly (after stated unit conversion)	$3.15/lb

Mark	Meaning	Example
Wrong	Source says something different	Min order 50 vs 500
Invented	Source says nothing	"Lead time: 2 weeks" when no lead time is given

Meets standard: zero invented values. A partner audit works best: they don't know what the table "should" say.

6. When to stop and use a spreadsheet

AI extracts and labels; **formulas calculate**. Totals, landed cost, sorting, and "cheapest" belong in a spreadsheet where every calculation is visible and repeatable. Keep the master copy in the spreadsheet, not in the chat. If the data is already structured (a CSV export), import it directly: a deterministic import is exact; AI adds risk for no benefit.

Worked example: three resin quotes

Prompt:

```
Extract Quote A (between the markers) into a table with exactly these
columns: Supplier | Unit price (USD/lb) | Minimum order (lb) |
Lead time (business days) | Freight terms | Payment terms |
Certifications stated | Source (quote + line).
If a value is not stated, leave the cell blank. If you convert units,
show the original in the Source column. Do not calculate totals or
recommend a supplier.
---[Quote A]---
```

Audit findings (first pass): Quote B lead time "10 days" (invented; B states none); Quote C "ISO 9001 certified" (invented; C says "certs available on request"); Quote B unit price $4.20/lb (correct conversion from $42 per 10-lb bundle, but the original must be shown).

Spreadsheet landed cost for 1,200 lb: A = $3,780 + freight (unknown, FOB origin); B = $5,040 including freight; C = $3,660 + $144 freight = $3,804. **Note to buyer:** "C is the lowest known landed cost; A is lower only if its freight is under $24. Ask A for a freight quote and B for a lead time before ordering."

Checklist: extraction you can trust

- ☐ Columns, units, and formats defined before extraction.
- ☐ Source column included.
- ☐ "Leave blank if not stated" in the prompt.
- ☐ One labeled document at a time; unneeded sensitive data stripped.
- ☐ Every cell audited by a partner: correct / wrong / invented.
- ☐ Zero invented values in the final table.
- ☐ Math done with spreadsheet formulas.

Do and don't

Do	Don't
Require blanks for missing values	Accept a "complete" table without asking where each value came from
Show unit conversions explicitly	Let the AI convert units silently
Use formulas for totals and comparisons	Ask the AI which supplier is cheapest
Check scans and PDFs against the original	Trust OCR on handwritten or faded forms
Strip bank details and names first	Paste whole email chains with signatures

Try this at work

1. Pick a comparison you do regularly (quotes, bids, schedules, policy versions).
2. Write the column list with units and a source column.
3. Extract two or three non-sensitive documents and have a coworker audit every cell.
4. Move the math into a spreadsheet and note the unknowns that need a phone call.

Glossary

Term	Meaning
Extraction	Pulling specific values from documents into structured fields
Invented value	A value the source doesn't state but the AI supplied
Cell audit	Checking every table cell against its source
Landed cost	Total cost delivered: price × quantity + freight and fees
OCR	Optical character recognition: reading text from images or scans
Source column	A column recording where each value came from

Key takeaways

- Define columns and require blanks.
- Audit every cell: zero invented values.
- AI extracts; spreadsheets calculate.