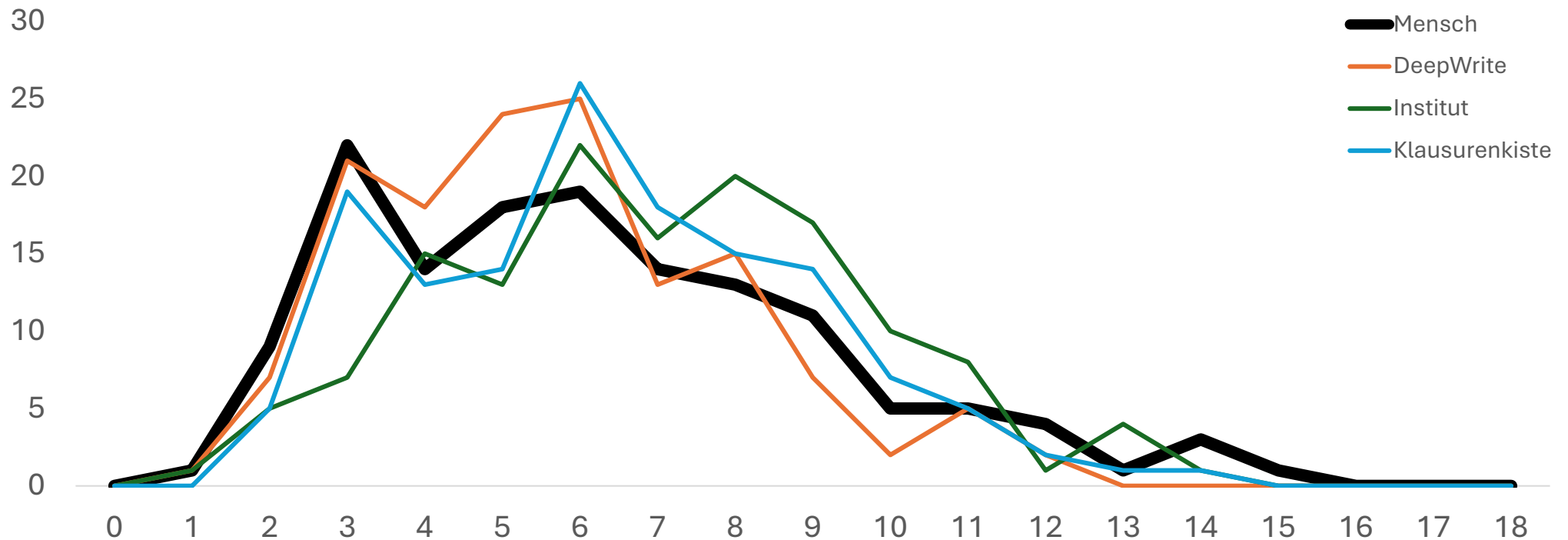


Vergleich: Notenverteilung bei KI-Korrektur und menschlicher Korrektur

	Mensch	DeepWrite	Institut	Klausurenkiste
Median	6	5	7	6
Durchschnitt	6,21	5,66	7,00	6,38
Prädikat	14%	6%	17%	11%
Nicht bestanden	23%	21%	9%	17%



Das Schaubild zeigt auf den ersten Blick, dass die menschliche Korrektur (dick dargestellt) nah an den KI-Korrektursystemen liegt. Die KI der Klausurenkiste und von DeepWrite neigen eher ins Mittelfeld, die KI des Instituts für Rechtsdidaktik hat eher im oberen Notenbereich eine Spitze.

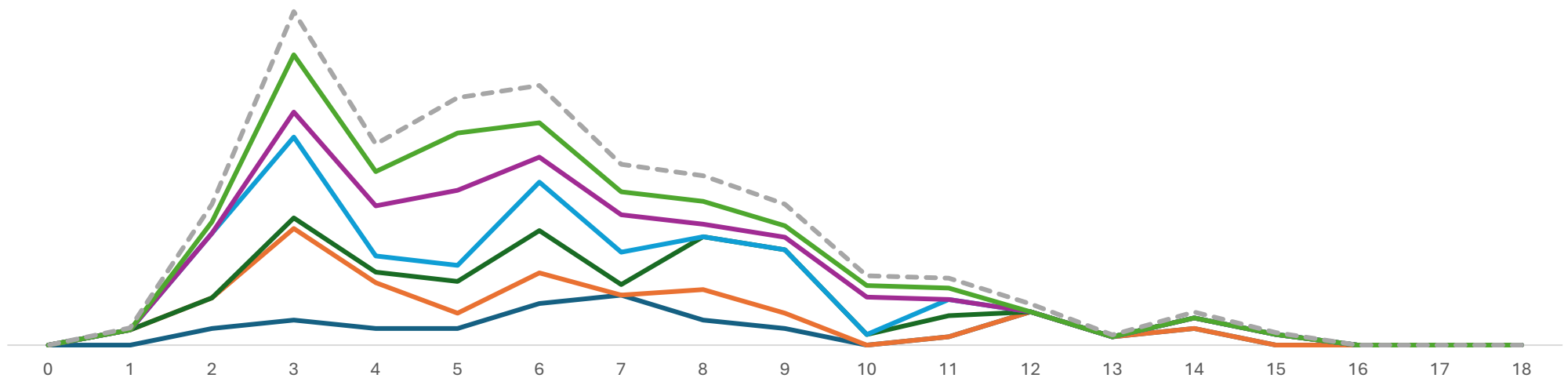
	Mensch	DeepWrite	Institut	Klausurenkiste
0	0	0	0	0
1	1	1	1	0
2	9	7	5	5
3	22	21	7	19
4	14	18	15	13
5	18	24	13	14
6	19	25	22	26
7	14	13	16	18
8	13	15	20	15
9	11	7	17	14
10	5	2	10	7
11	5	5	8	5
12	4	2	1	2
13	1	0	4	1
14	3	0	1	1
15	1	0	0	0
16	0	0	0	0
17	0	0	0	0
18	0	0	0	0

Auch eine Darstellung als „Heatmap“ zeigt, dass die Unterschiede (hier in absoluten Zahlen) überschaubar sind.

Aufschlüsselung nach Korrekturassistenten / Korrekturassistentinnen

Ein Blick auf die Korrekturassistent/innen zeigt, dass diese durchaus unterschiedlich bewertet haben – das mag aber auch an der verschiedenen Qualität der zugewiesenen Pakete liegen (jede Zeile stellt eine andere Korrekturperson dar; im Schaubild ist die gestrichelte Linie die Summe).

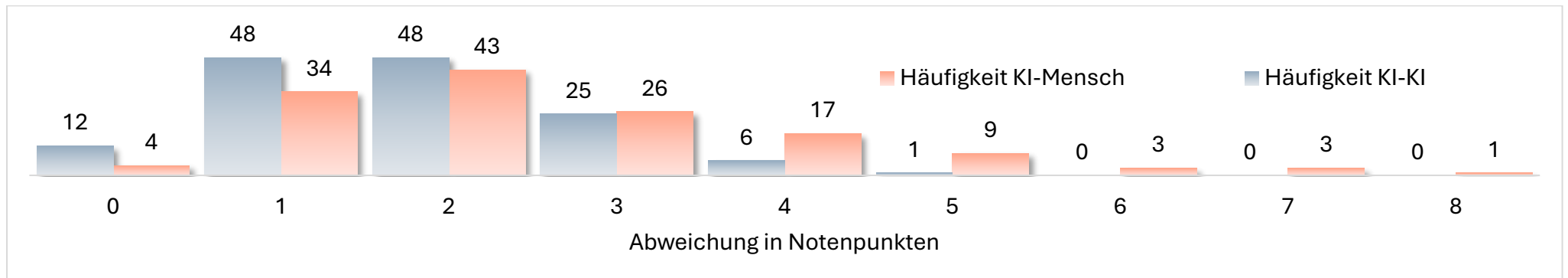
18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0%	0%	0%	0%	6%	3%	12%	3%	0%	6%	9%	18%	15%	6%	6%	9%	6%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	6%	11%	0%	11%	6%	17%	33%	11%	6%	0%
0%	0%	0%	4%	4%	0%	0%	8%	4%	23%	19%	4%	15%	12%	4%	4%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	6%	0%	0%	0%	12%	18%	6%	6%	29%	24%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	14%	5%	5%	14%	9%	27%	18%	9%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	4%	4%	4%	8%	8%	13%	21%	13%	21%	4%	0%	0%
0%	0%	0%	1%	2%	1%	3%	4%	4%	8%	9%	10%	14%	13%	10%	16%	6%	1%	0%



Umfang und Intensität der Abweichungen

Abweichungen der KI-Korrekturen zur menschlichen Korrektur und zur Korrektur anderer KI-Systeme bewegen sich überwiegend zwischen 0 und 3 Punkten (nach § 30 I 3 JAPO würde dies ohne Drittgutachten als Mittelwert gelten). Nur in 6 Fällen weichen KIs untereinander um 4 Punkte ab, in einem Fall um 5 Punkte. Bei den menschlichen Korrekturen gibt es **eine** Abweichung um 8 Punkte, 3 um 7 Punkte, 3 um 6 Punkte sowie 9 um 5 Punkte. Nur bei 4 Klausuren stimmen *alle* Bewertungen exakt überein. Dies zeigt leider, dass die Bewertung von Klausuren keine exakte Wissenschaft ist.

Differenz	Häufigkeit KI-Mensch	Häufigkeit KI-KI
8	1	0
7	3	0
6	3	0
5	9	1
4	17	6
3	26	25
2	43	48
1	34	48
0	4	12



Bewertungsunterschiede zwischen den eingesetzten KI-Systemen

Abweichung	DeepWrite	Klausurenkiste	Institut
-8	1	0	0
-7	0	2	0
-6	1	0	2
-5	2	2	1
-4	7	3	2
-3	12	12	3
-2	20	13	7
-1	20	16	29
0	39	28	18
+1	15	28	28
+2	16	18	21
+3	4	9	11
+4	2	2	9
+5	1	4	7
+6	0	2	0
+7	0	1	2
+8	0	0	0
Gesamt	-76	24	111

Die KI-Systeme haben sehr unterschiedliche Bewertungstendenzen:

- **DeepWrite** bewertet tendenziell strenger als die menschlichen Korrekturen – in einem Fall sogar 8 Punkte schlechter.
- Die **Klausurenkiste** kommt der menschlichen Korrektur am nächsten.
- Die **KI des Instituts für Rechtsdidaktik** bewertet tendenziell eher besser als die menschlichen Korrektoren – in zwei Fällen sogar 7 Punkte besser.

Im Rahmen der Auswertung haben wir uns auf Fälle fokussiert, in denen die maximale Abweichung am größten war. Diese (und einige weitere) Klausuren hat Prof. Harnos selbst korrigiert und bewertet:

- **Abweichung von 8 Punkten** – menschliche Korrektur: 14 Punkte; DeepWrite: 6 Punkte; Klausurenkiste: 7 Punkte; Institut: 9 Punkte; Prof. Harnos: Grenzfall zwischen 9 und 10 Punkten (**Befund:** fehlendes Gespür des menschlichen Korrektors für Schwerpunktsetzung und großzügige Rohpunktevergabe; strenge Rohpunktevergabe durch DeepWrite).
- **Abweichung von 7 Punkten** – menschliche Korrektur: 9 Punkte; DeepWrite: 3 Punkte; Klausurenkiste: 2 Punkte; Institut: 3 Punkte; Prof. Harnos: 3 Punkte (**Befund:** inhaltlich nicht nachvollziehbare Einträge in der Rohpunktetabelle des menschlichen Korrektors – Vergabe von Rohpunkten für nicht bearbeitete Probleme; nachvollziehbare KI-Bewertung).
- **Abweichung von 7 Punkten** – menschliche Korrektur: 3 Punkte; DeepWrite: 8 Punkte; Klausurenkiste: 10 Punkte; Institut: 10 Punkte; Prof. Harnos: 8 Punkte (**Befund:** menschliche Korrektur hat zutreffende Ausführungen, die *in überlangen Absätzen versteckt* waren, nicht positiv gewürdigt).
- **Abweichung von 6 Punkten** – menschliche Korrektur: 2 Punkte; DeepWrite: 6 Punkte; Klausurenkiste: 8 Punkte; Institut: 7 Punkte; Prof. Harnos: Grenzfall zwischen 5 und 6 Punkten (**Befund:** viel zu strenge Bewertung durch menschliche Korrektur).

Abweichungen nach Korrekturassistent/innen

Nicht nur Abweichungen der KI haben uns beschäftigt, sondern auch Abweichungen der Korrekturassistentinnen und -assistenten. Diese haben natürlich unterschiedliche Pakete, so dass eine Vergleichbarkeit nur bedingt möglich ist. Zudem haben wir geschaut, wie nah (oder entfernt) einzelne Korrekturpersonen von der Note der KI waren. Das Spektrum war hier bemerkenswert – von fast 3 Punkten im Schnitt bis zu weniger als einem Punkt im Schnitt:

Anzahl Klausuren	Prädikatsquote	Durchschnittsnote	Nichtbestehensquote	Median	KI-DifferenzMin	KI-Differenz-Durchschnitt	KI-Differenz-Max	Korrekturzeit	Anmerkungen (in Zeichen)
33	24%	7,42	15%	7	0	1,515151515	6	14:15 Min.	3.064,30
18	0%	4,28	50%	3,5	0	2,722222222	7	k.A.	k.A.
26	19%	8,00	4%	8	0	2,243589744	8	k.A.	k.A.
17	6%	4,41	53%	3	0	2,019607843	5	k.A.	k.A.
22	14%	6,00	9%	5	0	0,96969697	4	27:25 Min.	4.409,86
24	8%	5,50	25%	5	0	1,361111111	5	k.A.	k.A.